Title Page

Variation in Y chromosome, mitochondrial DNA and labels of identity in Ethiopia

Christopher Andrew Plaster

The Centre for Genetic Anthropology

Department of Genetics Evolution and Environment

University College London

Declaration of ownership

I, Christopher Andrew Plaster, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Statement of work performed by Christopher Plaster

Approximately 20% of all DNA extractions, 40% of all Y chromosome genotypings and 65% of mtDNA sequencing data was generated by myself, with the remainder generated by technicians and students (as part of other and ongoing projects at TCGA) prior to the commencement of my PhD studies. I checked all previous data generated by others for consistency with the data generated by myself. All processing of mtDNA data, and all statistical analysis was performed by myself. All DNA samples and ethnographic information was collected in the field by Dr Ayele Tarekegn.

Abstract

There is a paucity of genetic studies of Ethiopia. This thesis aims to establish the extent and distribution of variation in NRY and mtDNA genetic markers as well as ethnic and linguistic labels of identity in 45 ethnic groups. A wide range of NRY and mtDNA haplogroups were observed, including both those typically observed in Africa and those more frequently observed outside Africa. Significant correlations were revealed between NRY and mtDNA diversity. Nearly all ethnic groups were significantly differentiated from each other, although the pattern of similarities indicates some recent gene flow between northern ethnic groups and some groups to the south. Significant correlations were observed between almost all measures of linguistic and ethnic similarity and measures of NRY and mtDNA genetic distance. A wide range of values for diversity of sample donors' ethnic and linguistic identity was observed across groups. There was no evidence that the language of the sample donor is more likely to be inherited from parents of one sex rather than the other. There was a general decrease in the proportion of sample donors speaking the traditional language of their ethnic groups compared with the donor's parents and grandparents, with a corresponding increase in the proportion of sample donors speaking Amhara as a first language. Restricting analysis to samples from donors with ethnic identity in common with their parents and grandparents had significant effects on the degree of genetic distinctiveness of ethnic groups. Increasing the level of resolution of NRY and mtDNA haplotypes did not substantially alter the patterns of diversity and distance observed in and amongst ethnic groups. This thesis makes an important contribution to understanding the distribution of genetic diversity in Ethiopia.

Acknowledgements

I would like to thank all the individuals who donated samples, as well as those who helped collect them. I would particularly like to thank Dr Ayele Tarekegn, for spending so much time in the field collecting samples and recording data, and also for his invaluable knowledge while writing this thesis. He is the only man I know to have met and interviewed 10,000 Ethiopians (which has to be some kind of record).

I would also like to thank all the people at UCL who have made the time pass very quickly indeed. Some of the people I'd like to thank, in no particular order, include Krishna Veeramah, Adam Powell, Yuval Itan, Sarah Browning, Naser Ansari Pour, Abi Jones, Kate Ingram, Laura Horsfall, Bryony Jones, Lauren Johnson, Andrew Loh, Anna Texieira, Rosemary Ekong and many others. I'd also like to thank Maria Elevant, Ripudaman Bains, and Olivia Creemer for their extreme patience (with me) and dedication to a task (at times), in helping me check the questionnaires of a few thousand Ethiopians. I also appreciate it that they have kindly ignored that I had promised them dinner as payment for helping me out. Thanks again. I'd also like to thank all the undergraduate students I have supervised over the years for the comic relief that they have provided. In particular Rohan Wagle and Farzeen Rauf. The knowledge that one is now a commercial pilot, and the other will be a physician, is terrifying.

I would very much like to thank my sponsor, and my supervisors Neil Bradman and Mark Thomas, without whom none of this would have been possible. I'd also like to thank my family for their love and support they have given me. Lastly, I am deeply indebted to Michelle, for her continued love and support she has provided over the years, and especially over the past few months, as my own personal standards have dropped of late, and without whom I would by now be a particularly unwashed and undernourished individual.

Contents

Abbreviations

ASD	Averaged Squared Distance
CC1	Chartered City 1, Addis Ababa
CC2	Chartered City 2, Dire Dawa
DNA	Deoxyribonucleic Acid
ETPD	Exact Test of Population Differentiation
HVS1	Hyper-Variable Segment 1 of the mtDNA chromosome
K2P	Kimura 2 Parameter
KYA	Thousand Years Ago
LG	Linguistic group
MS	Microsatellite
MSV	Mean microsatellite repeat length variance
mtDNA	Mitochondrial DNA
MYA	Million Years Ago
NRY	Non Recombining portion of the Y-chromosome
PCO	Principle Coordinate analysis
PCR	Polymerase Chain Reaction
SNNP	Southern Nations Nationalities and Peoples province
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TCGA	The Centre for Genetic Anthropology
UEP	Unique Event Polymorphism
VSO	Variable Sites Only

Chapter1: Introduction

1.1 A brief history of Ethiopia

1.1.1 The geography of Ethiopia

Present day Ethiopia is a land-locked country located in eastern Africa, bordered by Eritrea to the north-east, by Sudan to north-west, by Kenya to the south, and to the east and south-east by Djibouti and Somalia respectively (Figure 1.1). The majority of the area of the country consists of mountainous terrain, with a highland plateau extending from the west with an elevation of over 1500m (Phillipson 1998; Zewde 2001). These highlands are bisected north-east to south-west by the Rift Valley, which extends from the Red Sea, creating the low lying desert and active volcanoes of the Danakil depression in the Afar region, passing through central Ethiopia and separating the northern and south-eastern highland regions, and continues south dotted with craters and lakes, ending with Lake Turkana (formerly Lake Rudolf) at Ethiopia's southern border with Kenya (Phillipson 1998; Zewde 2001) (Figure 1.1). Such rugged terrain results in a variety of climates and rainfall patterns over a relatively small geographic area (Phillipson 1998; Zewde 2001). The Danakil depression in the Afar region is one of the hottest and driest places on Earth, whereas the Simien Mountains in the northern Amhara region 250km to the west of Danakil often receives substantial snowfall (Pankhurst 1998; Phillipson 1998). The vast highland areas that dominate the country generally experience temperate conditions, despite Ethiopia lying between 3° and 15° north of the equator (Zewde 2001), whereas the more low lying areas experience climates more typical of its tropical latitude, with desert-like conditions in the east, dry scrubland in the south, and tropical broad-leaf forest found in the west of the country (Pankhurst 1998; Phillipson 1998; Zewde 2001).

Figure 1.1 Shaded relief map of contemporary Ethiopia (source: Perry-Castañeda Library Map Collection)

1.1.2 Early hominids in Ethiopia

According to Phillipson (1998), "Ethiopia can claim the longest archaeological record of any country in the world". Paleontological material uncovered in Ethiopia has also been extremely useful in uncovering the ancestry of anatomically modern humans, with the discovery of fossil hominid remains unearthed in the Middle Awash valley of Ethiopia spanning the past six million years (White et al. 2003). Perhaps the most famous example of fossil hominid remains uncovered in Ethiopia is the *Australopithecus afarensis* specimen known as "Lucy" (Johanson and White 1979), dated at 3.2 MYA, but much older hominid specimens have since been discovered, including the latest find named *Ardipithecus ramidus* dated at 4.4 MYA (White et al. 2009). Ancient remains of the most immediate ancestors of anatomically modern humans have been discovered in Ethiopia (White et al. 2003), with a pair of *Homo sapiens* specimens known as Omo I and Omo II dated at 195 KYA, which have the distinction of being the oldest anatomically modern human fossils yet discovered (McDougall et al. 2005). Ethiopia's proximity to Western Asia, the wealth of its fossil record over the period in which anatomically modern human evolved, and more recent genetic evidence, have led to the suggestion that Ethiopia is a prime candidate region from which anatomically modern humans first migrated out of Africa (Quintana-Murci et al. 1999; Stringer 2003; Kivisild et al. 2004; Ramachandran et al. 2005; Armitage et al. 2011)

1.1.3 Modern human history in Ethiopia

Little is known of the lifestyle specifics of pre-historic modern humans in Ethiopia prior to 1000BC (partly due to the hiatus in archaeological fieldwork in Ethiopia that occurred during the Derg period, and also due to the higher proportion of all archaeological work in the country to date concentrating on the period after 1000BC (David Phillipson, personal communication)), but undoubtedly a major transition would have been the move from a purely hunter-gatherer lifestyle to the use of domesticated plants and animals. Early farming took place in Ethiopia from at least the early first millennium BC (and possibly much earlier (Ehret 2002; Marshall and Hildebrand 2002)), with Ethiopia constituting a likely centre for domestication of both indigenous crops (*Teff*, *Enset*) as well as those which have since gone on to be cultivated worldwide (Coffee) (Phillipson 1998; Ehret 2002; Marshall and Hildebrand 2002). The beginnings of the use of domesticated plants and animals in Africa is

reviewed by Marshall and Hildebrand (2002), with the available evidence suggesting that animal husbandry generally predated the use of domesticated crops, but this observation could also be because of the lack of archaeological material, through the poor preservation conditions for seeds and cereals in the tropical regions of African, and an insufficient level of archaeobotanical research undertaken in the region. The earliest evidence for the use of domesticated animals in Ethiopia is dated to between 1500BC and 500BC at both the site of Gobedra in the northern highlands near Axum, and in the rift valley at Lake Beseka, and the use of domesticated cereals (*Teff*) is dated at 500BC at a site near Axum (Marshall and Hildebrand 2002).

As evidenced by epigraphic material, the early records of the modern human habitation of the lands of present day Ethiopia go back to at least to the third millennium BC through Egyptian accounts of trade with the 'Land of Punt', and the archaeological record including tools and pottery extends much further back in time (reviewed by Pankhurst (1998), Phillipson (1998) and Ehret (2002)). Extensive trade was undertaken in goods such as gold, ivory, salt, spices and slaves between groups in the lands of present day Ethiopia and neighbouring groups by routes overland and along the Red Sea, and there is some evidence of Ethiopian trade with countries as far afield as India and China (Pankhurst 1998; Phillipson 1998).

By the first millennium BC there is evidence of a strengthening of cultural links between the highlands of Ethiopia and the lands across the Red Sea, with the emergence of a culture known as D'mt, with stone architecture, sculpture, and writing appearing for the first time in Ethiopia, and of a kind which had previously only been seen in the southern Arabian peninsular (Phillipson 1998). This period may also have coincided with the introduction of early Ethiosemitic languages from southern Arabia (Kitchen et al. 2009). An advanced culture based around the town of Axum seems to have come to prominence by the beginning of the first millennium AD (Pankhurst 1998; Phillipson 1998; Zewde 2001; Ehret 2002). The Axumites maintained trade links with many cultures in the region including Roman North Africa and the Mediterranean, and spoke and wrote in the native Semitic language, Ge'ez. Axum is known to have adopted Christianity in the 4th century AD, and undertook many great building projects, including the massive stone stelae that still stand inside the contemporary town of

Axum (Pankhurst 1998; Phillipson 1998; Ehret 2002). The Axumite culture eventually went into decline, and by the 12th century AD control of the northern highlands was under the Cushitic-speaking Zagwe culture, which was notable for undertaking the construction of the famous rock-hewn churches found in Lalibela (Pankhurst 1998; Phillipson 1998; Ehret 2002). In the 13th century AD the Zagwe were eventually overthrown by the Semitic-speaking Amhara, who 're-established' the 'Solomonic' dynasty of emperors, professing that they could trace their ancestry to Menilek, son of King Solomon and the Queen of Sheba (Pankhurst 1998; Phillipson 1998; Ehret 2002).

Adherents of Islam were in contact with the Christian peoples of Ethiopia almost from its inception in the 7th century AD, and relations were mostly cordial due to their involvement in trade (Pankhurst 1998). By the end of the 15th century, the Islamic culture of the Adal Sultanate (based in what is today mainly the Somali and southern Afar regions) was threatening the Christian control of the northern highlands. The Portuguese had recently made contact with the Emperor of the Christian highlands through missionaries and diplomatic envoys, and the Ethiopian Emperor requested Portuguese aid to help defend against the Ottoman backed Adal incursions. The Portuguese sent troops to aid the Ethiopian defence, and successfully prevented any further advances of the Adal Sultanate. Years of conflict between the Islamic and Christian states had weakened them both, and from the mid 16th century AD the Cushitic speaking Oromo pastoralists were expanding their territory from their heartland in the low lying southern region of present day Ethiopia into the previous conflict regions of the southern and eastern highlands. This Oromo expansion was also accompanied by an apparent two-sided cultural assimilation, with instances of both the indigenous peoples and the immigrant Oromo adopting each other's ethnicities, and the practice of Oromo pastoralism changing to a sedentary agricultural lifestyle (Pankhurst 1998; Ehret 2002).

In the latter part of the 19th century the Ethiopian southern border expanded to its furthest extent to encompass the highlands and lowlands to the south, east and west of the country, lands that had not previously been part of the Ethiopian state, but for many centuries had been linked to the northern empire in both trade and conflict (Pankhurst 1998; Zewde 2001). The cultures that were now brought under imperial control

included: the Semitic speaking Gurage in the region south of present day Addis Ababa, who allied themselves briefly with the Cushitic speaking Hadiya to try to prevent the northern advance, the Semitic speaking Harari who were incorporated to the south-east in what is now the Somali region, the kingdom of the Omotic speaking Wolayta in the southern highlands which was conquered after much resistance from the native population as well as from the nearby Omotic peoples of the Dawuro and Konta, and the Omotic speaking Kefa kingdom which was finally subjugated after its long refusal to become a vassal state of the empire to its north-east, which was only achieved with the aid of the previously conquered Dawuro, Konta and the peoples of the present day Gambella region (Zewde 2001). By the end of this campaign of expansion the southern borders of Ethiopia resembled those of the present day, the capital city of Addis Ababa was fairly well-established, and Menilek who was previously the King of the Shewa province in the central highland region and responsible for much of the territorial expansion, had become emperor of all Ethiopia (Zewde 2001).

By the end of the 19th century, having previously been mostly ignored by European nations in the 'scramble for Africa', Ethiopia found itself in conflict with Italian interests in expansion of its territories in East Africa (Pankhurst 1998; Zewde 2001). In 1889 a treaty was signed between Ethiopia and Italy, in which the Ethiopian version granted Italy territory in what is now Eritrea, but in the Italian version, unbeknown to the Ethiopians, substantial concessions on Ethiopian sovereignty were also granted (Pankhurst 1998; Zewde 2001). This ultimately led to conflict, and in 1896 during the Adwa campaign in northern Ethiopia, the Italian force was defeated by a multi-ethnic Ethiopian force, preventing any further Italian military incursions into Ethiopian territory for almost forty years (Pankhurst 1998; Zewde 2001). In 1935 Italy successfully invaded and occupied Ethiopia, uniting their territory of Eritrea in the north with Italian-held Somalia in the south (Pankhurst 1998; Zewde 2001). The Italians were removed with the aid of British forces who were at war with Fascist Italy in 1941, and Ethiopia was eventually returned to Imperial rule (Pankhurst 1998; Zewde 2001).

In 1974, following on from a recent famine, and a perceived inability of the Emperor and his government to modernise the country, the military overthrew the government and installed a communist junta known as the *Derg* (Pankhurst 1998; Zewde 2001).

The *Derg* undertook land reform, but were ultimately deeply unpopular after their bloody purges and forced resettlements, and were eventually overthrown by a coalition of rebel forces known as the Ethiopian People's Revolutionary Democratic Front (EPRDF) in 1991 (Pankhurst 1998; Zewde 2001). In 1994 a constitution had been adopted, and by 1995 Ethiopia's first multi-party elections had been held (https://www.cia.gov/library/publications/the-world-factbook/geos/et.html).

Figure 1.2 Languages of Ethiopia (from www.ethnologue.com)

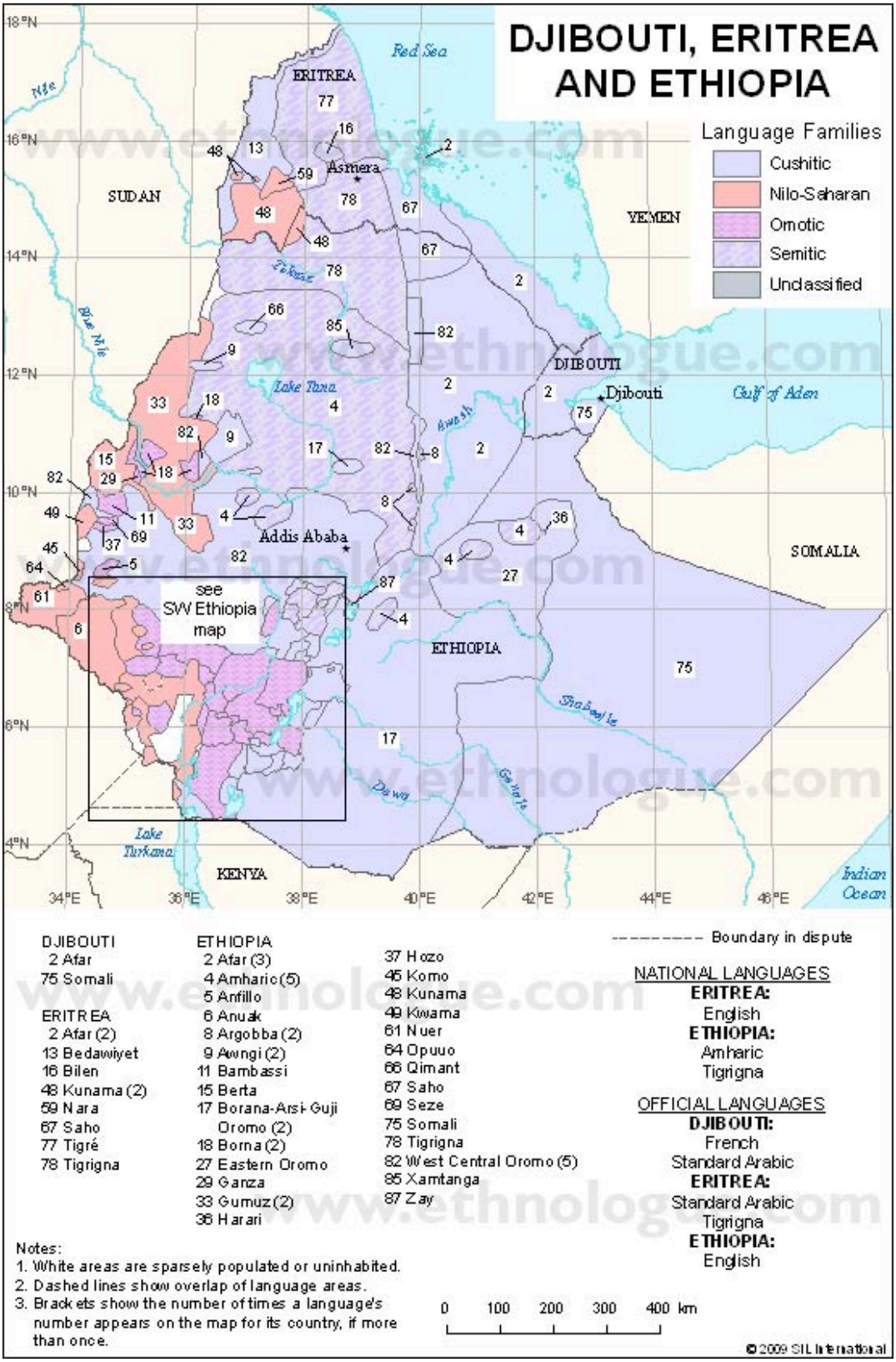
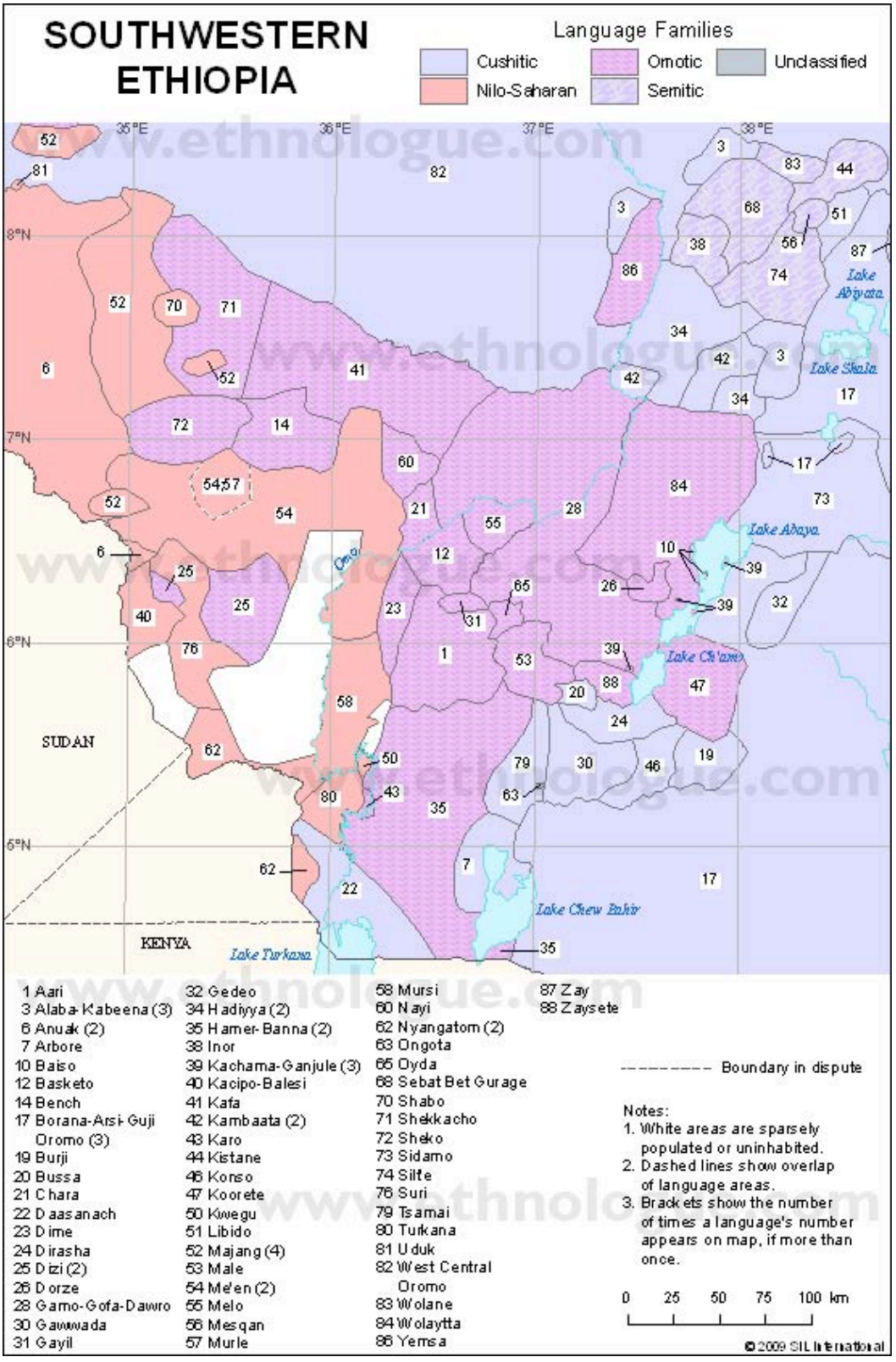


Figure 1.3 Languages in south-west Ethiopia (from www.ethnologue.com)



1.2 The contemporary peoples of Ethiopia

Ethiopia today contains over 80 ethnic groups and spoken languages (Ethiopian Population Census Commission 2007; https://www.ethnologue.com/). It has a population of over 85 million (2010 estimate (https://www.cia.gov/library/publications/the-world-factbook/geos/et.html)), with the Oromo and Amhara ethnic groups constituting 34.5% and 26.9% of the total population respectively, and Somali, Tigray, Sidama, Gurage, Wolayta, Hadiya, Afar, Gamo and Gedoe with census sizes over 1 million (Ethiopian Population Census Commission 2007). The remaining 11.3% consists of the vast majority of ethnic groups, with most comprising less than 0.1% each of the total population.

Languages spoken in Ethiopia belong to two main language families Afro-Asiatic and Nilo-Saharan, of which Afro-Asiatic comprises of three main groups: Semitic, Cushitic and Omotic. Speakers of Cushitic languages are widespread in Ethiopia, with representatives in both the highlands and low lying regions (Figure 1.2). Semitic speakers can primarily be found in the northern highland regions, whereas Nilo-Saharan speakers are predominantly located in the west of the country (Figure 1.2). The Omotic group of languages are entirely restricted to Ethiopia (www.ethnologue.com), with the highest incidence to be found in the south-west of the country (Figure 1.3). The following ethnic groups were included in this study (with collection locations indicated in section 2.4):

1.2.1 Afar

The Afar are an ethnic group primarily inhabiting the Afar Administrative Region in North-East Ethiopia. The Afar region is generally a low-lying desert, and is bordered by the Somali region to the south, Eritrea and Djibouti to the north and east, and the Amhara and Tigray regions to the West. The majority of the samples included in this study were collected from the towns of Asaita and Dubti. According to the 2007 Ethiopian census, the Afar ethnic group numbered 1,276,374 people. The Afar language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.2 Agaw

The Agaw are an ethnic group primarily inhabiting the Amhara Administrative Region in the northern highlands of Ethiopia. The Agaw can be found living in several geographically separated populations across a large area primarily inhabited by the Amhara ethnic group (www.ethnologue.com). The Agaw included in this study were collected from both a western population known as the Agaw Awi, that were mainly collected in the town of Gimja Bet Maryam, and an eastern population known as the Agaw Hemra, that were mainly collected from the towns of Sekota, Tsitska and the surrounding countryside. According to the 2007 Ethiopian census, the Agaw Awi ethnic group number 631,565 people, and the Agaw Hemra ethnic group numbered 267,851 people. The Agaw language belongs to the Central Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.3 Alaba

The Alaba are an ethnic group primarily found in the Southern Nations Nationalities and Peoples (SNNP) Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Alaba are the Kembata to the west, the Hadiya to the south and north-west, the Gurage to the north, and the Oromo to the east (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Alaba Kulito and its environs. According to the 2007 Ethiopian census, the Alaba ethnic group numbered 233,299 people. The Alaba language belongs to the east Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.4 Amhara

The Amhara are the second largest ethnic group in Ethiopia, and according to the 2007 Ethiopian census numbered 19,870,651 people. The majority of Amhara people are found in the northern highland regions in the Amhara Administrative Region, which is bordered by the Afar Region to the east, the Tigray Region to the north, Sudan to the west and the Oromia Region to the south. The majority of samples included in this study were collected in Addis Ababa, Dese and Mekane Selam. The Amhara language belongs to the North Semitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.5 Anuak

The Anuak are an ethnic group primarily found in the Gambella Administrative Region, in western Ethiopia, and also in neighbouring Sudan. The neighbouring ethnic groups to the Anuak are the Nuer to the north-west, the Oromo to the north and the Mejenger to the east. They are bordered by Sudan to the south-west (www.ethnologue.com). The

majority of the samples included in this study were collected from the towns of Gambela, Gog and Itang and their environs. According to the 2007 Ethiopian census, the Anuak ethnic group numbered 85,909 people. The Anuak language belongs to the Nilo-Saharan language family (www.ethnologue.com).

1.2.6 Ari

The Ari are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Ari are the Hamer and the Bena to the south, the Maale to the east, the Basketo to the north, and the Dime to the west (www.ethnologue.com). The majority of the samples included in this study were collected in the towns of Bako Gazer district and Jinka town and its environs. According to the 2007 Ethiopian census, the Ari ethnic group numbered 290,453 people. The Ari language belongs to the south Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.7 Basketo

The Basketo are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Basketo are the Gofa to the north and east, the Ari to the south and the Dime to the west (www.ethnologue.com). The majority of the samples included in this study were collected from the district capital town of Laska and its environs. According to the 2007 Ethiopian census, the Basketo ethnic group numbered 78,284 people. The Basketo language belongs to the north Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.8 Bena

The Bena are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to Bena are the Hamer to the south, the Mursi to the west, the Tsemay to the east, and the Ari to the north (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Key Afer and its environs. According to the 2007 Ethiopian census, the Bena ethnic group numbered 27,022 people. The Bena language belongs to the south Omotic branch of the Afro-Asiatic language family.

1.2.9 Bench

The Bench are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Bench are the Me'en to the south and east, the Kefa to the north, and the Sheko to the west (www.ethnologue.com).

The majority of the samples included in this study were collected from the towns of Aman and Mizan Teferi. According to the 2007 Ethiopian census, the Bench ethnic group numbered 353,526 people. The Bench language belongs to the north Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.10 Burji

The Burji are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Burji are the Konso to the west, the Kore to the North, and the Oromo to the south and east (www.ethnologue.com). The majority of the samples included in this study were collected from the village of Berek and the town of Soyema. According to the 2007 Ethiopian census, the Burji ethnic group numbered 71,871 people. The Burji language belongs to the east Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.11 Busa

The Busa are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Busa are the Zayse to the north-east, the Dirasha to the south-east, the Gawada to the south and the Maale to the north-west (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Gidole and its environs. According to the 2007 Ethiopian census, the Busa ethnic group numbered 10,458 people. The Busa language belongs to the east Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.12 Dasanach

The Dasanach are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Dasanach are the Nyangatom to the north-west, the Turkana to the south and the Hamer to the north-east. They are bordered to the south by Kenya (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Omorate and its environs. According to the 2007 Ethiopian census, the Dasanach ethnic group numbered 48,067 people. The Dasanach language belongs to the east Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.13 Dawuro

The Dawuro are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Dawuro are the Wolayta

to the east, the Gamo, the Gofa and the Melo to the south, the Kefa to the west and the Oromo to the north (www.ethnologue.com). The Manja ethnic group live amongst the Dawuro, particularly in the town of Waka and its environs where the majority of the samples included in this study were collected (Ayele Tarekegn personal communication). According to the 2007 Ethiopian census, the Dawuro ethnic group numbered 543,148 people. The Dawuro language belongs to Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.14 Dirasha

The Dirasha are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Dirasha are the Kore to the east, the Burji, Konso, and Gawada to the south, the Busa to the west, and the Zayse to the north (www.ethnologue.com). The majority of samples included in this study were collected from the towns of Arba Minch and Gidole. According to the 2007 Ethiopian census, the Dirasha ethnic group numbered 30,081 people. The Dirasha language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.15 Dizi

The Dizi are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Dizi are the Me'en to the north, the Suri to the west and south, and the Mursi to the east across the Omo National Park. The majority of samples included in this study were collected from the town of Maji and the village of Tum. According to the 2007 Ethiopian census, the Dizi ethnic group numbered 36,380 people. The Dizi language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.16 Dorze

The Dorze are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Dorze are the Wolayta to the north and east, and the Gamo to the south and west (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Chencha and Dorze. There was no entry for Dorze in the 2007 Ethiopian census, but according to the 1994 census, the "Dorzie" were 28,990 people. The Dorze language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.17 Gamo

The Gamo are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Gamo are the Dorze to the north, and the Gofa to the west and south (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Arba Minch and Mirab Abaya. According to the 2007 Ethiopian census, the Gamo ethnic group numbered 1,107,163 people. The Gamo language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.18 Ganjule

The Ganjule are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Ganjule are the Gamo to the north and east, and the Zayse to the south (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Shele Mela. Formerly, the Ganjule lived on an island in Lake Chamo, but have now mostly relocated to the town of Shele Mela (www.ethnologue.com). There was no entry for Ganjule in the 2007 Ethiopian census, but according to the 1994 census, the Ganjule were 1,146 people. The Ganjule language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.19 Gedeo

The Gedeo are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Gedeo are the Sidama to the north, and the Oromo to the east, south and west (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Dila and Wenago. According to the 2007 Ethiopian census, the Gedeo ethnic group numbered 986,977 people. The Gedeo language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.20 Genta

The Genta are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Genta are the Ganjule to the south-east and the Gamo to the north and west (Ayele Tarekegn personal communication). The size of the Genta population is unknown, as the Genta ethnic group was not listed in either the 1994 or the 2007 Ethiopian census. The majority of the samples included in this study were collected from the towns of Arba Minch and Shele Mela, and also from the Genta Meche Peasant Association. The Genta language

belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.21 Gewada

The Gewada are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Gewada are the Konso to the east, the Dirasha to the north, the Tsemay to the west, and the Oromo to the south (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Gewada. According to the 2007 Ethiopian census, the Gewada ethnic group numbered 68,600 people. The Gewada language belongs to the East Cushitc branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.22 Gobeze

The Gobeze are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Gobeze are the Mashile to the south-east, the Busa to the north-east, the Dirasha to the north and the Gewada to the west (Ayele Tarekegn personal communication). The majority of the samples included in this study were collected from the administrative section (or *kebele*) of Dega Mashile and the town of Gidole. The size of the Gobeze population is unknown, as the Gobeze ethnic group was not listed in either the 1994 or the 2007 Ethiopian census. The Gobeze language belongs to the Cushitic branch of the Afro-Asiatic language family, and is considered a dialect of Busa (www.ethnologue.com).

1.2.23 Gofa

The Gofa are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Gofa are the Gamo on the east, the Basketo on the west, the Ari on the south, and the Wolayta and the Dawuro on the north (www.ethnologue.com). The majority of samples included in this study were collected from the town of Arba Minch and its environs. According to the 2007 Ethiopian census, the Gofa ethnic group numbered 363,009 people. The Gofa language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.24 Gurage

The Gurage are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Gurage are the Hadiya and Alaba to the south, the Yem to the west, and the Oromo to the north and east (www.ethnologue.com). The majority of the samples included in this study were

collected from the towns of Buta Jira, Indibir and from Addis Ababa. According to the 2007 Ethiopian census, the Gurage ethnic group numbered 1,867,377 people. The Gurage people speak a variety of related dialects, all of which belong to the South Semitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.25 Hadiya

The Hadiya are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Hadiya are the Gurage to the north, the Alaba and Kembata to the east, the Wolayta to the south, and the Yem to the west (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Hosaina. According to the 2007 Ethiopian census, the Hadiya ethnic group numbered 1,284,373 people. The Hadiya language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.26 Hamer

The Hamer are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Hamer are the Bena to the north, the Arbore to the east, the Dasanach to the south, and the Nyangatom to the south west (www.ethnologue.com). The majority of the samples that were included in this study were collected from the town of Dimeka. According to the 2007 Ethiopian census, the Hamer ethnic group numbered 46,532 people. The Hamer language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.27 Kefa

The Kefa are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring groups to the Kefa are the Oromo to the north, the Dawuro to the east, the Shekecho to the west, and the Bench, Chara, Me'en, and Nao to the south (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Bonga and Shishinda. According to the 2007 Ethiopian census, the Kefa ethnic group numbered 870,213 people. The Kefa language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.28 Kembata

The Kembata are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Kembata are the Wolayta to the south, the Hadiya to the west and north, and the Alaba to the east (www.ethnologue.com). The majority of the samples included in this study were

collected from the town of Durame. According to the 2007 Ethiopian census, the Kembata ethnic group numbered 630,236 people. The Kembata language belongs to East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.29 Konso

The Konso are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Konso are the Oromo to the south, the Burji to the east, the Dirasha to the north, and the Gewada to the west (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Arba Minch, Karat, and Konso. According to the 2007 Ethiopian census, the Konso ethnic group numbered 250,430 people. The Konso language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.30 Konta

The Konta are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Konta are the Dawuro to the north, east and south, with the Kefa to the west. The majority of the samples included in this study were collected from the town of Chida. According to the 2007 Ethiopian census, the Konta ethnic group numbered 83,607 people. The Konta language belongs to the Omotic branch of the Afro-Asiatic language family, and is considered a dialect of Dawuro (www.ethnologue.com).

1.2.31 Kore

The Kore are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Kore are the Oromo to the north and east and the Burji, Konso and Dirasha to the south. They are bordered by the shores of Laka Chamo and Lake Abaya to the west (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Arba Minch and its environs and also the villages of Berek and Buniti. According to the 2007 Ethiopian census, the Kore ethnic group numbered 156,983 people. The Kore language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.32 Maale

The Maale are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Maale are the Gamo and Gofa to the north, the Busa to the east, the Gewada, Tsemay and Bena to the south, and

the Ari to the west (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Jinka in Bako Gazer District. According to the 2007 Ethiopian census, the Maale ethnic group numbered 98,114 people. The Maale language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.33 Mashile

The Mashile are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Mashile are the Gobeze to the north-west, the Busa to the north-east and south, the Dirasha to the north and the Gewada to the west. The majority of the samples included in this study were collected from the administrative sections of Dega Mashile and Qola Mashile, and also from the town of Gidole. The size of the Mashile population is unknown, as the Mashile ethnic group was not listed in either the 1994 or the 2007 Ethiopian census. The Mashile language belongs to the Cushitic branch of the Afro-Asiatic language family, and is considered a dialect of Busa (www.ethnologue.com).

1.2.34 Mejenger

The Mejenger are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Mejenger are the Anuak to the west in the Gambella region, the Oromo to the north, the Shekecho to the east, and the Sheko to the south (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Indris and Kokobe. According to the 2007 Ethiopian census, the Majenger ethnic group numbered 21,959 people. The Majenger language belongs to the Nilo-Saharan language family (www.ethnologue.com).

1.2.35 Nuer

The Nuer are an ethnic group primarily found in the Gambella Administrative Region, in western Ethiopia, and neighbouring Sudan. The neighbouring ethnic groups to the Nuer are the Anuak and the Oromo to the east. They are bordered by Sudan to the north, west and south (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Akobo and Gambella. According to the 2007 Ethiopian census, the Nuer ethnic group numbered 147,672 people. The Nuer language belongs to Nilo-Saharan language family (www.ethnologue.com).

1.2.36 Oromo

The Oromo are the largest ethnic group in Ethiopia, with a population of 25,489,024 according to the 2007 Ethiopian census. The Oromo are widespread in Ethiopia and northern Kenya, with a large population in the Oromo Administrative Region that stretches across the middle of Ethiopia, bordering most of the other Administrative Regions. The majority of the samples included in this study were collected from Addis Ababa and Jimma. The Oromo language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.37 Shekecho

The Shekecho are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Shekecho are the Majenger to the west, the Sheko to the south, the Kefa to the east, and the Oromo to the north (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Bonga and Rimichi. According to the 2007 Ethiopian census, the Shekecho ethnic group numbered 77,678 people. The Shekecho language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.38 Sheko

The Sheko are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Sheko are the Majenger and Shekecho to the north, the Anuak to the west, the Me'en to the south, and the Bench to the east (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Bonga and Tepi. According to the 2007 Ethiopian census, the Sheko ethnic group numbered 37,573 people. The Sheko language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.39 Sidama

The Sidama are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Sidama are the Wolayta to the west, the Gedeo to the south, and the Oromo to the east and north (www.ethnologue.com). The majority of the samples included in this study were collected from the town Yirga Alem in the district of Daale. According to the 2007 Ethiopian census, the Sidama ethnic group numbered 2,966,474 people. The Sidama language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.40 Somali

The Somali are an ethnic group found in East Africa, predominantly in Somalia, but also with large populations in Ethiopia, Djibouti and north-east Kenya (www.ethnologue.com). The Somali ethnic group in Ethiopia is primarily found in the Somali Administrative Region, and is bordered by the Afar region to the north and the Oromia region to the west. They border Somalia from the north-east to the south-west. The majority of the samples included in this study were collected from the town of Jijiga. According to the 2007 Ethiopian census, the Somali ethnic population in Ethiopia numbered 4,581,794 people. The Somali language belongs to the East Cushitic branch of the Afro-Asiatic language family.

1.2.41 Tigray

The Tigray are an ethnic group primarily found in the Tigray Administrative Region, which borders the Amhara region to the south, the Afar Region to the east, Eritrea to the north and Sudan to the west. The Tigray samples used in this study consist of incidental collections made while recruiting from other ethnic groups, with the largest collection of samples from the town of Sekota. According to the 2007 Ethiopian census, the Tigray ethnic population in Ethiopia numbered 4,483,892. The Tigray language belongs to the north Semitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.42 Tsemay

The Tsemay are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Tsemay are the Bena and Hamer to the west, the Arbore to the south, the Gewada to the east, and the Maale to the north (www.ethnologue.com). The majority of the samples included in this study were collected from the town Key Afer and the administrative section (*kebele*) of Enchete. According to the 2007 Ethiopian census, the Tsemay ethnic group numbered 20,046 people. The Tsemay language belongs to the East Cushitic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.43 Wolayta

The Wolayta are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Wolayta are the Dawuro to the west, the Gamo and Dorze to the south, the Sidama to the east, and the Hadiya and Kembata to the north (www.ethnologue.com). The majority of the samples included in this study were collected from the towns of Arba Minch and Wolayta Sodo.

According to the 2007 Ethiopian census, the Wolayta ethnic group numbered 1,707,079 people. The Wolayta language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.44 Yem

The Yem are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Yem are the Oromo to the west, north and south, and the Hadiya and Gurage to the east (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Fofa. According to the 2007 Ethiopian census, the Yem ethnic group numbered 160,447 people. The Yem language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.2.45 Zayse

The Zayse are an ethnic group primarily found in the SNNP Administrative Region, in south-west Ethiopia. The neighbouring ethnic groups to the Zayse are the Gamo to the north, the Ganjule to the east on the shores of Lake Chamo, the Dirasha to the south, and the Busa and Maale to the west (www.ethnologue.com). The majority of the samples included in this study were collected from the town of Arba Minch and its environs, and the village of Elgo Dombe. According to the 2007 Ethiopian census, the Zayse ethnic group numbered 17,884 people. The Zayse language belongs to the Omotic branch of the Afro-Asiatic language family (www.ethnologue.com).

1.3 Previous work on the distribution of human genetic variation amongst the contemporary peoples of Ethiopia

From the earliest studies of Ethiopian genetics it has been known that the peoples of the region are diverse and distinct from many other groups surveyed in sub-Saharan Africa. In *History and Geography of Human Genes*, Cavalli-Sforza et al. (1994) using 'classical markers' showed that the Ethiopian populations in their study appeared to occupy an intermediate position between West African groups and North Africans in a principal component analysis, with admixture between sub-Saharan and non-sub-Saharan Africans suggested as a possible explanation.

The intermediate position of some Ethiopians populations between sub-Saharan Africa and Eurasia was further supported by Wilson et al. (2001), who in applying STRUCTURE analysis of autosomal microsatellites observed that 62% of Oromo and Amhara speaking Ethiopians fall into an ancestral cluster that included Ashkenazi Jews, Armenians and Norwegians while only 24% of Ethiopians belonged to a cluster containing South African Bantu speakers and Afro-Caribbeans. More recently Tishkoff et al. (2009) using STRUCTURE analysis that utilised 1327 markers and surveyed 121 African populations, showed that Ethiopians of the Burji and Konso ethnic groups, collected in Kenya (but in Ethiopia are more typically found in the southern SNNP region), were observed to have 71% and 73% respectively of their ancestry in common with clusters dominated by other East African Cushitic speaking groups.

To date however, the greatest number of population genetics studies utilising samples collected from Ethiopian populations have analysed sex-specific inherited markers, in particular markers on the paternally inherited non-recombining portion of the Y chromosome (NRY) and the maternally inherited mitochondrial genome (mtDNA). The utility of these markers in current population genetic studies is reviewed in Underhill and Kivisild (2007), and I will go over some of the main points for two sex-specific genetic systems below. The two main benefits in using NRY and mtDNA markers over autosomal markers is firstly their lack of recombination. This allows for more easily recoverable phylogenies than is possible for the autosomal markers, allowing for the easier identification of geographically restricted clades, which could be indicative of past historical migrations (although caution is advised when inferring demographic history from the geographic pattern of clades, see for example the critique of one such method by Panchal and Beaumont (2010)). The second benefit in using the sex-specific systems is their much smaller effective population size relative to autosomal markers due to their haploid mode of inheritance through one sex only. This means that such systems are more prone to genetic drift, and therefore more sensitive to changes in the demography, resulting in more conspicuous genetic structure (Jobling et al. 2004), but this also means that such systems are also particularly sensitive to ascertainment bias (although the sample collection strategy used in this thesis attempts to avoid this problem, see Chapter 2). Additionally, in comparing data from the male lineage (NRY) to the female lineage (mtDNA), unique insights into sex-specific demographic history can be gained when the pattern of variation in the two systems differs (see for example

Seielstad et al. (1998) for the evidence for a higher female migration rate, and the evidence of sex-specific differences in the expansion of Bantu speakers observed by Wood et al (2005)). The continuing work on the discovery of new NRY binary polymorphism by many different groups has led to several refinements of the Y chromosome phylogenetic tree and the use of various nomenclature systems over the years. The Y Chromosome Consortium (YCC 2002) presented a phylogeny of 153 haplogroups, and a systematic nomenclature system, with the main clades designated by letters A-R, approximately defining the deepest rooted to the most derived clades respectively. The phylogeny and nomenclature system was later revised again by Karafet et al. (2008) due to the discovery of new binary markers, resulting in a phylogeny of 311 haplogroups. This nomenclature system added two new main clades (S and T), as well as other re-arrangements in the overall topology, and I have used this nomenclature system in this thesis. Unlike the NRY phylogeny and nomenclature system, the mtDNA phylogeny is regularly redrawn as more data become available, and the nomenclature system is still evolving, although attempts are being made to collate all available phylogenetic data (van Oven and Kayser 2009). As the naming of the main mtDNA clades in the early studies was done in the order that the clades were discovered, this has resulted in the deepest rooted (and in most cases Africa specific) clades being designated with the letter L, unlike the NRY system which uses A (Jobling et al. 2004). The high mutation rate in the mtDNA relative to autosomal systems, and the limited size of the mtDNA genome, results in high levels of homoplasy and difficulties in tree construction. These problems however are being overcome with by the increasing availability of data on the whole mtDNA genomes (see Behar et al. (2008) for an example) resulting in more robust phylogenies, and with the identification of panels of easily genotyped markers that are associated with specific clades (Behar et al. 2007).

One of the earliest studies of both NRY and mtDNA variation in Ethiopians was conducted by Passarino et al. (1998), and utilised Ethiopian samples collected from 77 patients and staff in a hospital in Addis Ababa. Despite the relatively small number of NRY markers used in this study, it was apparent that the Ethiopian samples were quite distinct from the Senegalese samples also included in the study for comparative purposes. Notably, the derived state of the sY81 marker was not observed in any of the Ethiopian samples, but was modal in the Senegalese samples, the derived YAP marker

was observed in almost all the Senegalese samples (98.9%), but only 50.0% of Ethiopians, and the derived state of the p12f2 marker was observed at 27.8% frequency in the Ethiopians in the study, but was absent from the Senegalese. In their RFLP assaying of the Ethiopian mtDNA to detect known haplogroups, it was observed that 20.3% of the samples were of haplogroup M, a haplogroup the authors note as an "Asian" type, and at that time had not been observed in any other "Caucasoid" or "African" lineages, and that its high frequency in Ethiopians is indicative of either "interchanges with Asians" or "was present in the ancient Ethiopian population and was carried by groups who migrated out of Africa".

The most comprehensive survey of NRY haplogroups in Ethiopian populations to date was undertaken by Semino et al. (2002), which also utilised data from an earlier study of NRY haplogroups present in global populations by Underhill et al. (2000). In this study, 48 Amhara and 78 Oromo samples were assayed for a panel of NRY haplogroup markers identified by Underhill et al. (2001), and the results compared with Ethiopian and Khoisan samples from the Underhill et al. (2000) study. The authors noted that the clades with the deepest divisions in the Y chromosome phylogeny (groups I and II according to the Underhill et al. (2000) nomenclature used by the authors, Haplogroups A and B according to nomenclature of Karafet et al. (2008)) although of different types, occurred in both the Ethiopian groups surveyed and the Khoisan from the Underhill et al. (2000) study, but were not observed in the Senegalese samples included for comparative purposes. It was observed that over 70% of the Ethiopian samples and almost all the Senegalese samples belonged to the group III (the E clade according to the Karafet et al. (2008) nomenclature), of which 80.6% of the Senegalese and none of the Ethiopian samples had the derived state for the sY81 marker (listed as the synonym M2 in the paper, defining the clade E1b1a according to Karafet et al. (2008) nomenclature). Interestingly, the authors noted a significant difference ($p<0.005$) between Amhara and Oromo samples in the frequency of the derived state of the M35 marker (defining the E1b1b1 clade according to Karafet et al. (2008)), with frequencies of 35.4% and 62.8% observed in each group respectively. A significant difference ($p<0.0001$) between the two ethnic groups was also observed in the frequencies of the derived state of the p12f2 marker, with frequencies of 33.4% and 3.8% observed in the Amhara and Oromo respectively. The authors hypothesised that the presence of the derived state of the p12f2 marker as well as the derived state of the M70 marker

(denoting the T clade according to Karafet et al. (2008) nomenclature) in the Ethiopian groups, could be explained by back migrations of peoples from Asia.

In a later study by Semino, et al. (2004) using the same Amhara and Oromo samples as in the Semino, et al. (2002) study, as well as the set of global population samples from the earlier work, investigated distribution of subtypes of NRY haplogroups E and J. Overall, it was observed that 79.5% of the Oromo and 45.8% of the Amhara belong to the E clade, of which the highest frequency subclade for both groups was E1b1b1a (according to Karafet et al. (2008) nomenclature, defined by the derived state of M78 in the study), and was observed at frequencies of 35.9% and 22.9% of samples in the Oromo and Amhara respectively. Using the Karafet et al. (2008) nomenclature, other clades observed in both ethnic groups were E1b1* (defined by the derived state of marker P2, observed at 12.8% of Oromo and 10.4% of Amhara), E1b1b1* (defined by the derived state of M35, observed in 19.2% of Oromo 10.4% of Amhara) and E1b1b1c (defined by the derived state of marker M123, observed at 5.1% of Oromo and 2.1% of Amhara. The clades E1b1c, E1b1b1d and E2 (indicated by the derived states of markers M329, M281 and M75 respectively) were only observed in the Oromo, and at less than 3% frequency. As noted previously in the Semino, et al. (2002) study, the frequency of haplogroup J (as defined by the derived state of the p12f2 marker), was far higher in the Amhara (35.4%) than the Oromo (3.8%). In the Semino, et al. (2004) study it is shown that the majority of the Amhara that belong to haplogroup J (33.3% of samples) have the derived state of the M267 marker (defining the J1 clade according to Karafet et al. (2008) nomenclature), and this was also observed in the Oromo at 2.6% frequency.

In a study by Cruciani, et al. (2004), 12 samples of the Wolayta ethnic group were surveyed for haplogroups present in the E clade, along with 34 Amhara samples, 25 Oromo and 12 samples of "mixed Ethiopians". They observed that clades E*(xE1b1b) (indicated by the ancestral state of marker M215, derived state of marker SRY4064), E1b1b1*, E1b1b1a, E1b1b1c1 and E1b1b1e (indicated by the derived states of markers M35, M78, M34 and V6 respectively) were present in all four groups, with the exception of clade E1b1b1c1 which was not observed in the mixed Ethiopian sample set. In this study, several global populations were also surveyed for haplogroups in the

E clade (including several European populations, North Africans, Near Eastern groups, and other sub-Saharan African groups), and the authors note that haplogroup E1b1b1e was only observed in the Ethiopians and in a single sample from Somalia and Kenya. In a later study by Cruciani, et al. (2007) the Ethiopian samples from the earlier study were assayed for further markers in the E1b1b1a (derived M78 marker) clade, but in this instance the 25 Ethiopian Oromo samples were combined with 7 Kenyan (Borana) Oromo samples. It was observed that in the Amhara and combined Ethiopian and Kenyan Oromo, all samples that were haplogroup E1b1b1a (8.8% and 40.6% respectively) were of the more derived clade E1b1b1a1b (denoted by the derived state of marker V32). Of the E1b1b1a samples in the Wolayta and mixed Ethiopian samples set (16.7% and 33.3% respectively), 8.3% of the Wolayta samples and 25% of the mixed Ethiopians were of haplogroup E1b1b1a3 (denoted by the derived state of marker V22), with the remaining samples belonging to the E1b1b1a1b clade.

Of studies on the variation in mtDNA haplogroups present in Ethiopian ethnic groups, the most comprehensive survey to date was performed by Kivisild et al. (2004). In this study, samples from 53 Tigray (from both Ethiopia and Eritrea), 120 Amhara, 33 Oromo, 21 Gurage, 16 Afar and 28 samples of various Ethiopian ethnicities were screened for mtDNA coding region markers and sequenced for HVS1 and II variation, along with 115 Yemeni samples collected in Kuwait included for comparative purposes. The authors note that 52.2% of Ethiopian haplogroups and 45.7% of the Yemeni samples were of clades L0-L6, of which the remainder were sub-clades of haplogroups M and N. The authors contrast these frequencies with data from 416 Mozambican samples used by Pereira, et al. (2001) and Salas, et al. (2002), where almost all haplogroups were of clades L0-L6. Interestingly, regarding the L0-L6 haplotypes present in the Yemeni samples, the authors note that there is greater similarity with those found in Mozambique than those found in Ethiopia, with 49% of Yemeni L0-L6 haplotypes with a match in Mozambique, versus 9% with a match in Ethiopia. The authors surmise that, given the very similar overall frequency of L0-L6 clades in Ethiopia and Yemen, but the lack of substantial haplotype sharing between the two areas, both groups have been subject to “major demic influences from different sources-which they do not necessarily share”. Overall, the most common mtDNA clade observed was L3 and it and its subclades, accounted for 34.0% of samples, of which haplogroup M1 accounted for 17.0% of samples. L3(xM,N) was observed to be at

highest frequency in the Oromo, accounting for 27.3% of samples, and was observed at lowest frequency in the Tigray at 11.3%. The M1 clade was observed at highest frequency in the Oromo at 18.2%, and lowest in the Afar at 6.3%. Sub-clades of haplogroup N account for 30.7% of samples, of which (preHV)1 accounted for 10.4% of samples. The authors state that the observed frequencies of (preHV)1, along with haplogroup M1, were almost equal among Cushitic (Afar and Oromo) and Semitic (Tigray, Amhara and Gurage) speaking groups. The authors note that they observe significant differences in the frequencies of derived N lineages between their highest incidence in the Tigray (47.2%) to their lowest in the Afar (18.8%), which the authors state is consistent with the proximity of the Tigrinya region to the Red Sea coast and its potential source for gene flow with southern Arabia.

A study by Poloni, et al. (2009) utilised samples from south-western Ethiopia, namely the Nyangatom and Dasanach, which were assayed for mtDNA haplogroup markers and sequenced for the HVS1 and II variation, and then analysed alongside data from the earlier study by Kivisild, et al. (2004) and data from Tishkoff, et al. (2007) on samples collected from populations in Tanzania. In this study it was observed that approximately 95% of Nyangatom and Dasanach samples were of the L series of clades, with only about 5% of samples belonging to the M or N clades, which contrasts with the haplogroup frequencies observed in the more northern Ethiopian groups analysed by Kivisild et al. (2004). On genetic distance, the Nyangatom and Dasanach were observed to be significantly differentiated from both each other as well as from the Ethiopian ethnic groups in the Kivisild et al. (2004) study, when using both haplogroup and haplotype (HVS1) data. Interestingly, when the Nyangatom and Dasanach were compared with the Tanzanian populations from the Tishkoff et al. (2007) study, along with the northern Ethiopian groups from the Kivisild et al. (2004) study, it was observed that the northern Ethiopian groups cluster together in a multidimensional scaling plot of haplogroup level Φ_{st} distances, with the Nyangatom and Dasanach clustering with the Tanzanian groups. The Nyangatom and Dasanach, despite being significantly differentiated from each other, and from the northern Ethiopian groups, were observed to be not significantly differentiated from particular Tanzanian groups, with the Nyangatom not differentiated from the Turu, and the Dasanach not differentiated from the Turu, Datog and Burunge. The authors state that one purpose of their study was to see to what extent cultural and linguistic similarity follows genetic similarity. The

authors note that they did not observe a correlation between genetic distance and either geographic distance or linguistic distance, and this is highlighted by the apparent greater genetic similarity of the Nyangatom and Dasanach with Tanzanian groups despite over one thousand kilometres separating them, and also with the Nilo-Saharan speaking Nyangatom demonstrating greater similarity with the Niger-Congo speaking Turu as well the Cushitic speaking Dasanach demonstrating similarity with Niger-Congo, Nilo-Saharan and Cushitic speaking Tanzanian groups. The authors state that a possible explanation for this pattern of genetic diversity and distance amongst the pastoralist groups of southern Ethiopia and Tanzania could be due to periods high mobility and gene flow amongst these ethnic groups, followed by periods of isolation and genetic drift, as well as potentially recent ethnogenesis.

1.4 Aims

This thesis reports an initial study designed to establish the extent and distribution of human genetic variation in the sex specific genetic systems (non recombining portion of the Y chromosome (NRY) and the mitochondrial genome (mtDNA)) among 45 Ethiopian ethnic groups, widely distributed throughout the country and speaking Semitic, Cushitic, Omotic and Nilo-Saharan languages. Specifically I report:

- Levels of diversity in NRY and mtDNA.
- The distribution of identified diversity.
- The extent to which inferences are affected by the selection of batteries of markers used to characterise the systems.
- Evidence of gene flow and its directions.
- Implications of sampling strategies based on the declared ancestry of DNA donors.
- Correlations of genetic and social labels (identity, language and birthplace).

- Comparisons of inferences drawn from NRY and mtDNA and alternative measures of assessing diversity.
- An observed ratio of NRY and mtDNA diversity at the level of assessment used in this study.
- Variation in observed patterns in this study from those reported in West African groups.

It is anticipated that the results reported in this study will:

- Inform decision making in the selection of samples for high density genotyping and whole genome sequencing.
- Form a foundation for future detailed studies of the demographic histories of the peoples of Ethiopia.

Chapter 2: Materials and methods

2.1 Sample collection

2.1.1 Collection strategy

Buccal swab samples were collected from anonymous male donors over 18 years of age, unrelated at the grandparental level who gave informed consent. Towns and villages, as well as rural areas which were known to contain large numbers of a particular ethnic group, were targeted for sample collection. Samples were collected on a first come first taken basis, provided that donors were not related at the paternal grandparental level to any previous sample donor. Along with a buccal swab, each donor was questioned for information on their self declared ethnicity, languages spoken and place of birth with similar information collected about their parents, paternal grandfather and maternal grandmother. All sample collection, and completion of questionnaires (on behalf of the sample donor), was either undertaken by or under direct supervision of Dr Ayele Tarekegn (AT).

2.1.2 Buccal swab DNA sample collection

Donors were instructed on the procedure for buccal swab DNA collection after informed consent was given. The cotton swab is rubbed on the inside of both cheeks and the inside of the mouth of the donor for at least 20 seconds. Swabbing was either performed by AT, or by the sample donor themselves under supervision of AT. All samples were collected in duplicate. After swabbing is complete, the swab is immediately placed inside a sample collection tube containing 1ml solution of 0.05M Ethylenediaminetetraacetic acid (EDTA) and 0.5% Sodium Dodecyl Sulfate (SDS), that immerses the cotton end of the swab, and retards breakdown of the DNA sample during transit to the laboratory. Each sample collection tube is then securely sealed with PVC tape, and kept in as cool and dark place as possible until it can be transported to the laboratory.

2.1.3 Collection of ethnographic information on sample donors

For each sample donor, a questionnaire is completed in English by AT on behalf of the sample donor. Each anonymous donor was allocated an alphanumeric code that was written on the sample collection tubes of the donor's DNA sample and recorded on the donor's questionnaire. For each donor, the following information was recorded by AT:

-The donor's age

-Self declared cultural identity (or ethnicity)

- First and second language spoken

-The donor's religion

-Place of birth

-The cultural identity (or ethnicity) of the donor's father, mother, paternal grandfather and maternal grandmother

-First and second language spoken by the donor's father, mother, paternal grandfather and maternal grandmother

-The place of birth and current residence of the donor's father, mother, paternal grandfather and maternal grandmother

2.1.4 Non Ethiopian samples used for comparative purposes

In order to place variation amongst the Ethiopian ethnic groups in context, NRY and mtDNA data on four other geographically widely distributed groups in the TCGA collection were analysed and compared with the Ethiopian data:

-Igbo, collected in Calabar, Nigeria (IGB, 95 samples)

-Greek-Cypriots, collected in Nicosia, Cyprus (CYP, 126 samples)

-Fars, collected in Tehran, Iran (IRN, 92 samples)

-Halfawi, collected in Wadi Halfa, Northern Sudan (SUD, 60 samples)

NRY and mtDNA genetic data was generated by previous members of TCGA, using the same methods used in this thesis. Due to the poor quality of the Halfawi samples, comprehensive data was only available for the six NRY MS1 STRs and for 228bp (nucleotide position 16027 to 16254) of mtDNA HVS1. Despite the poor quality of these samples compared with the Ethiopian dataset and the three other additional groups, it was important to include in the analysis samples from a country (Sudan) where many of the ethnic groups found in Ethiopia are also known to reside. Consequently, all comparative analysis of the Ethiopian groups and the four additional

groups was performed using the variation in these six NRY MS1 STRs and 228bp of mtDNA HVS1.

2.2 Laboratory methods

2.2.1 Extraction of DNA from buccal swab samples

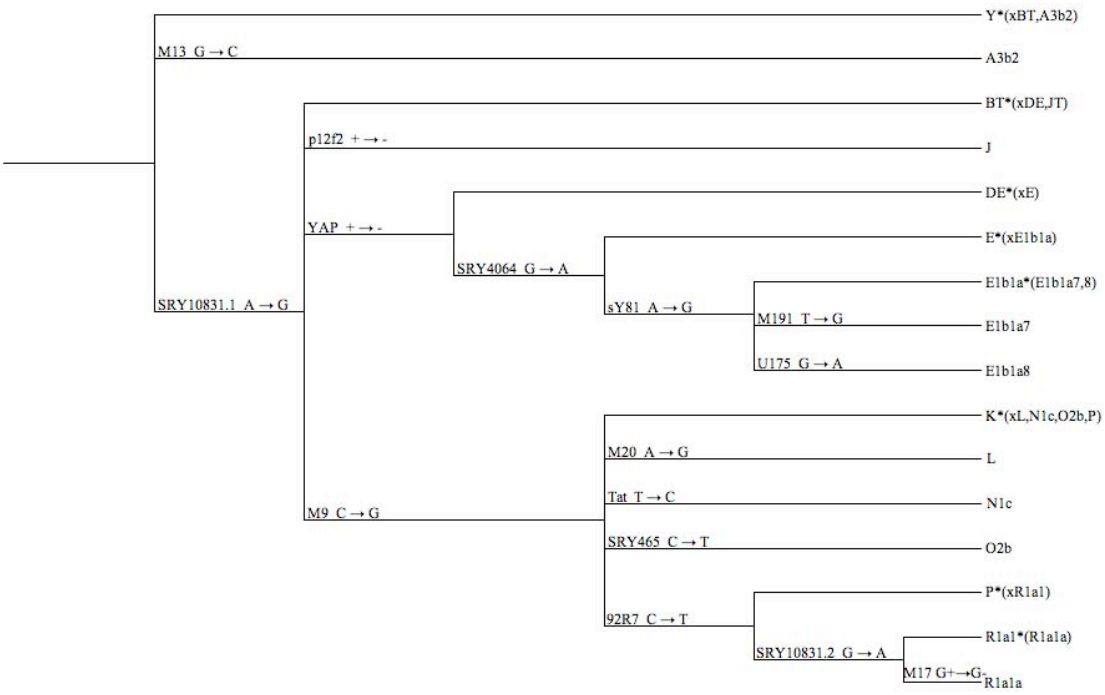
DNA from buccal swab samples was extracted using a phenol/chloroform method. To each buccal swab sample collection tube containing the swab and 1ml of 0.05M EDTA and 0.5% SDS, 0.8ml of 0.02 $mgml^{-1}$ of proteinase K solution is added, and then incubated at 56°C for two hours. 0.8ml of the digested buccal swab sample solution is then added to microfuge tubes containing a 1:1 mix of phenol/chloroform. The sample solution and phenol/chloroform is mixed, and then centrifuged for 10 minutes at over 2000g. The aqueous upper layer is then transferred to a microfuge tube containing 0.6ml of chloroform and 30µl of 5M NaCl. The sample solution and chloroform is mixed and centrifuged for 10 minutes at over 2000g. The aqueous upper layer is then transferred to a microfuge tube containing 0.7ml of chloroform. The sample solution and chloroform is mixed, and then centrifuged for 10 minutes at over 2000g. The aqueous upper layer is then transferred to a 2ml screw-top microfuge tube (used for long-term DNA storage) containing 0.7ml isopropanol. The sample and isopropanol solution is mixed, chilled at -20°C for two hours, and then centrifuged for 13 minutes at over 2000g. The supernatent is decanted off carefully to avoid dislodging the precipitated DNA pellet, and the microfuge tube allowed to dry for 1 minute while inverted at a 45° angle. 0.8ml of 70% ethanol is then added to the microfuge tube, and then centrifuged for a final 10 minutes at over 2000g. The supernatent is decanted off carefully, and the microfuge tube allowed to dry for 20 minutes while inverted at a 45° angle. The DNA is finally eluted by the addition of 400µl of TE (pH 8.0), and then mixed and incubated for 10 minutes at 56°C, briefly centrifuged, and then stored upright at -20°C until use.

2.2.2 Generation of data on NRY variation

Data on NRY variation in all samples was generated using the methods described in Thomas, et al. (1999). These methods characterise six STRs (DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393) and eleven biallelic UEP markers (92R7, M9, M13, M17, M20, SRY+465, SRY4064, SRY10831, sY81, Tat, YAP). STR sizes were determined according to nomenclature of Kayser et al. (1997). Samples with the

derived state of the SRY10831 marker, and the ancestral state of the YAP and M9 markers were further genotyped for the p12f2 marker using the methods described by Rosser et al. (2000). Samples with the derived state of the sY81 marker were further genotyped for markers M191 and U175 by Naser Ansari Pour, according to the method described in Veeramah et al. (2010). NRY clades were defined according to the nomenclature used by Karafet et al. (2008) (Figure 2.1).

Figure 2.1 Genealogical relationship of haplogroups defined by UEP markers according to the nomenclature of Karafet et al. (2008)



2.2.3 Generation of data on mtDNA HVS1 variation

Data on mtDNA HVS1 variation was obtained by sequencing using the methods described by Thomas, et al. (2002) with the following modifications:

-primer conL1-mod was replaced by primer conL849 (CTA TCT CCC TAA TTG AAA ACA AAA TA)

-primer conL2 was replaced by primer conL884 (TGT CCT TGT AGT ATA A)

-primer conH3 was replaced by primer conHmt3 (CCA GAT GTC GGA TAC AGT TC)

All samples with a minimum sequence covering nucleotides 16019-16400 were compared to the Cambridge Reference Sequence (Anderson et al. 1981) in order to

identify the polymorphic nucleotide positions (hereafter referred to as Variable Sites Only, VSO), with VSO haplotypes consisting of the nucleotide positions where substitutions, insertions or deletions occurred as well as the actual base change

2.2.4 Assays performed by others outside of TCGA

As it was not feasible to assay all the 5,756 Ethiopian samples for additional markers during the time allowed for my thesis studies, more extensive genotyping was performed on a subset of samples that had undergone whole genome amplification. This sample subset (377 samples, hereafter referred to as Ethiopian Ascertainment samples) was selected and whole genome amplified by Dr Sarah Browning of TCGA, during the period of my PhD studies, in an initial attempt to assemble a sample set on which to assess the extent of human genetic diversity in the peoples of the country. Having regard to the archaeological evidence of contact and possible demographic influence from the Arabian peninsula in the north-east, the general geographic distribution of religious affiliations and spoken language groups, the northern location of the Axumite empire in ancient times, and also the late 19th century conquest of present day southern Ethiopia by the Emperor Menilek, five ethnic groups representing a rough north-east to south-west transect were selected:

-Afar (AF, mainly Muslim Cushitic speakers, 74 samples)

-Amhara (AM, mainly Christian Semitic speakers, 76 samples)

-Anuak (AN, mainly Traditional Religion Nilo-Saharan speakers, 76 samples)

-Maale (ML, mainly Traditional Religion Omotic speakers, 75 samples)

-Oromo (OR, mixed but mainly Christian and Muslim, with some Traditional Religion, Cushitic speakers, 76 samples)

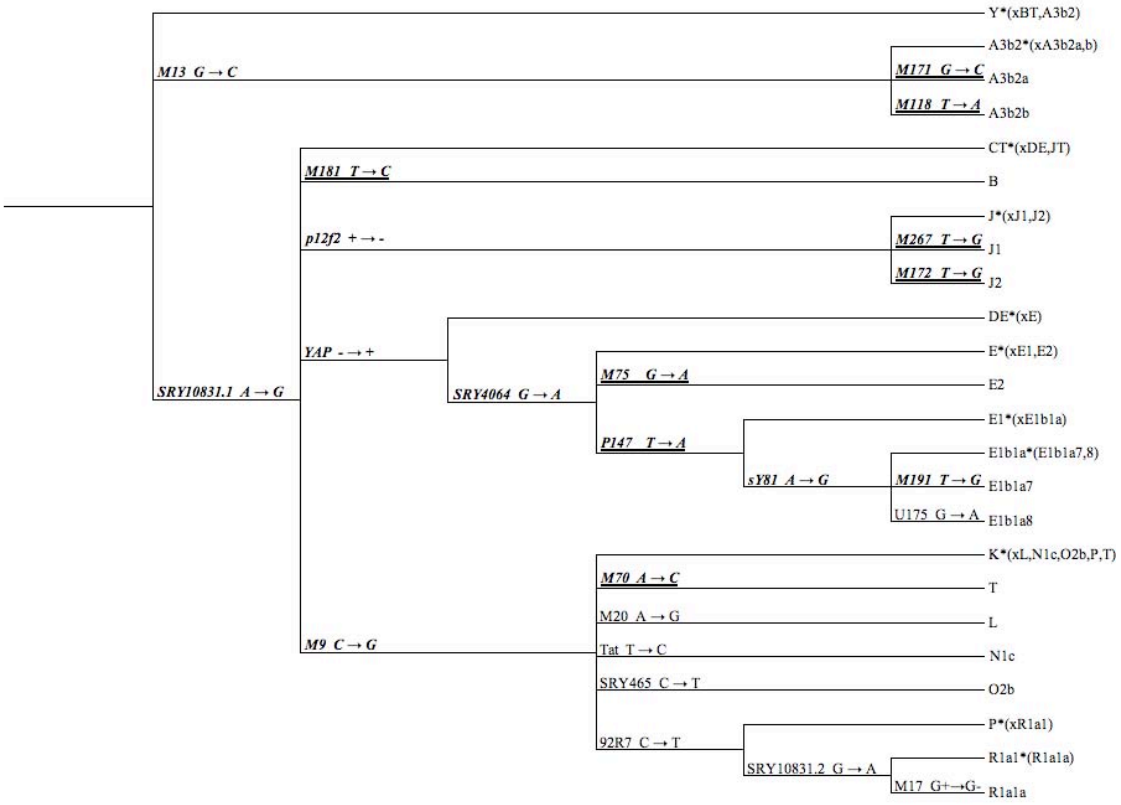
2.2.4.1 Genotyping of additional NRY haplogroup markers

Ethiopian ascertainment samples were submitted to Kbioscience UK (Hoddesdon, Hertfordshire, UK) for genotyping of additional NRY haplogroup markers after the genotyping of UEP and STR markers detailed in section 2.2.2 was completed. Genotyping assays were designed by Kbioscience using the SNP information and flanking sequences that I provided. The SNPs and flanking sequences are detailed in Table 2.1, with the relative genealogical position of the defined clades shown in Figure 2.2.

Table 2.1 NRY SNPs and flanking sequences submitted to Kbioscience for genotyping

Marker	Clade defined	SNP and submitted flanking sequence (SNP indicated by square brackets)
M171	A3b2a	GAAATCTGCTTTTTGTTTTGCAGAGAGCTTGGAGATAATTCTGGTGGCTGTGTGGAGTATGTGTTGGAGGTGAGTTGCTAGCTGAAGAATTAAAACAATAGTTTTAGCAGTTTGGGTAAGAGATGTTTACAGAAATGTTTTGTG[S]AATAAAACTGAACAGTCAGAGACCTATGAGATT
M118	A3b2b	ATTCTAAGTTTCACTTCCTGATCCACCACAGAAATCACTTTACAATGTTCTTCCCTTCCTCCATCACTGCATTCTTCTCAACCAGCTGACACTTGTGTTTTCTTTATA[W]GAGTAAGTGGTATCTTTCTTTTGTTAGTAAAGTTTATCTCAGAAGCTCCTATGGTAAAAGCAGCAGTAACCAAAGCAGAAGTTTCACATTAAAAGAAAACAAAGTTGTTGTCCTTAATTTCAAGGGAATCAGCACATGGTAGCT
M181	B	GCACTGGCGTTCATCATCTGGGAGCAGCTCAAAAGCCTCTCGCTCAGCCTCCGTGACGCCCTGGGGGTGTTCAACCCACATATACTGTAAAGACTAGGAGTAGGGTTGTGGACACCCCACCTCAGCCAACACTGAGCCCTGATGTGGACTCAACCTTGTAAGGAAAGCTGTAGAGAAATTGGAAGAAAAAATATAAACACATACAGACTCTGTCTTTACATTTCAAAATGCATGACTTAAAG[-/T]ATCAGGCA
P147	E1	CCCTACAAGGACTGGGCAAGTAATGCTAACTTTAATCAATTGATGAATAC[W]CCAGGAAGAGAACTTTTGGGGAACTAATGAGTTAATTGTTCATCATATAT
M75	E2	AAAGCCAAAACAGATTTCTAATGTACTGTGAAAAGACAATTATCAAACCACATCC[R]TATATATACAGAGAAATACCTTTATAAGAATAAAAATTCACAAATGCCTC
M267	J1	TTATCCTGAGCCGTTGTCCCTGTGTTTCCATTTCTCTTTTCCTCATTTCTCATCATCTACATTTCTCCTGTACTTGTTCATTAAATAATGATTCCTTGGATATACCAAGTCTGGATAGCGGATTCGATGGAAGCATTTTTGTAAATA[K]ACGTTCAGTATTTTGTGTGGAAGAACACAATCTAGCTGATGCCTGCAATCCCAGCCCTTTGGAAAGCGAGGTGGGTGGATTGCTTGAAGCTACGAGTTTGACACT
M172	J2	TTGAAGTTACTTTTATAATCTAATGCTTAATCTCTTTAAATATTTAAAATTAGGAGCCAGATGACCAGGATGCCCCAGATGAGCATGAGCCCTCTCCATCAGAAGATGCCCCATTATATCCTCATTCACCTGCCTCTCAGTATCAACAGGTAAAAAGGATTTTTCATTTTTATCCCCCAAACCCATTTTGATGCTT[K]ACTTAAAAGGTCTTCAATTATTATTTTCTTAAATATTTTGAAAGTCCAAACTTTCT
M70	T	GGTTATCATAGCCCACTATACTTTGGACTCATGTCTCCATGAGA[M]CTAAGACTACCACAACAGAATCCCTATAGTCCAGCCCTCAGATCACATACATGTACAGGCATGTTGAAGTAGTCGGACTTGAAGGAATCAGCCATTTCACCAAAACTCTGCAAACTGTACTCCTGGGTAGCCTGTTCAAATCCAAAAGCTTCAGGAGGCTGTTTACACTCCTGAAATAAAATATATTTCAGCAAGACAAAGGGAATAA

Figure 2.2 Genealogical relationship of haplogroups defined by UEP and the additional NRY haplogroup markers according to the nomenclature of Karafet et al. (2008). Markers with derived states in the Ethiopian ascertainment samples are indicated in italic type, additional haplogroup markers are italic and underlined.



2.2.4.2 Assaying of additional NRY STRs

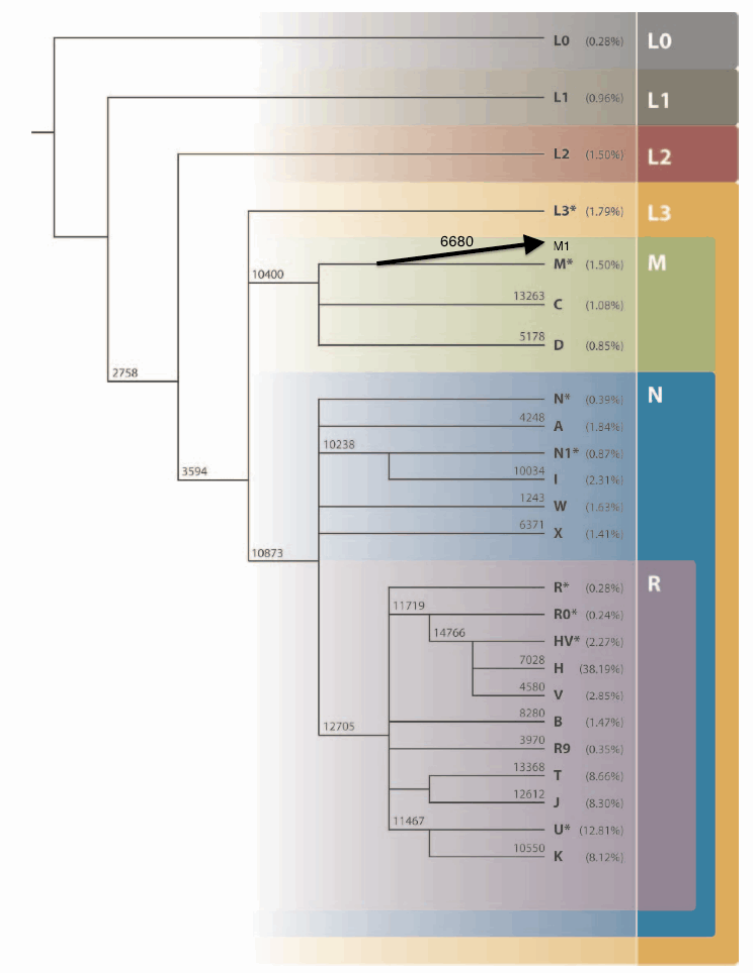
Ethiopian ascertainment samples were submitted to The Genome Centre (Barts and The London, Queen Mary University of London, London, UK) for assaying of additional NRY STR markers after the genotyping of UEP and STR markers detailed in section 2.2.2 was completed. The AmpFℓSTR Yfiler PCR amplification kit (Applied Biosystems) was used by The Genome Centre to determine the STR repeat size of 9 additional loci (DYS389I, DYS389II, DYS437, DYS438, DYS439, DYS448, DYS456, DYS635, Y GATA H4), as well as for 5 STR loci that overlap with those assayed using the methods described by Thomas et al. (1999) (DYS19, DYS390, DYS391, DYS392, DYS393). The AmpFℓSTR Yfiler kit also assays the DYS385 and DYS458 loci, but the information from these markers were excluded from the analysis as they were either a duplicated locus with similar repeat sizes (DYS385), so haplotype sharing could not easily be determined, or have a high incidence of partial repeat sizes (DYS458), so

preventing straight forward estimations of Rst and MSV. These 14 STRs were combined with the results for the DYS388 locus from section 2.2.2 in all analyses.

2.2.4.3 Genotyping of mtDNA haplogroup markers

Ethiopian ascertainment samples were submitted to Kbioscience UK (Hoddesdon, Hertfordshire, UK) for genotyping of mtDNA haplogroup markers after the sequencing of mtDNA HVS1 detailed in section 2.2.3 was completed. Genotyping was performed on the 22 coding region markers in the panel described in Behar, et al. (2007) with an additional marker for the T>C SNP at mtDNA position 6680 for the M1 clade (Figure 2.3). Haplogroup data was combined with HVS1 sequence data from section 2.2.3 to determine levels of homoplasy and diversity within clades.

Figure 2.3 Figure 4 from Behar, et al. (2007) showing the 22 markers and the clades they infer, with the phylogenetic location of the M1 SNP at mtDNA position 6680



2.3 Statistical Analysis

2.3.1 Genetic diversity metrics

2.3.1.1 Gene diversity (*h*)

Gene diversity, (*h*, the probability of randomly sampling two different alleles from a population), and its variance was estimated using the unbiased formulae of Nei (1987), which is defined as $h = n(1 - \Sigma\chi_i^2)/(n-1)$, where n is the sample size and χ_i^2 is the squared frequency of allele *i*th allele. The standard deviation (s.d.) is given as the square root of the variance.

2.3.1.2 Mean microsatellite variance (MSV)

The mean microsatellite variance (hereafter referred to as MSV), is the mean of the STR repeat size variances over all STR loci, and was recently used by Chiaroni et al. (2010) and Balaresque et al. (2010).

2.3.1.3 Dating of the STR variation in NRY clades

Dating of the STR variation in NRY clades used the method of Zhivotovsky et al. (2004) with the modification of Sengupta et al. (2006). This method uses the average of the squared differences (ASD) (Thomas et al. 1998) between the repeat sizes of each STR locus and the ancestral repeat size. This is expected to be μt for single step mutations, where μ is the mutation rate, and t is the time since the population diverged in generations. ASD was determined using the median value of the repeat size at each NRY STR locus (Sengupta et al. 2006). For a clade, the age of variation was estimated as the average of the per locus ages of variation. Two different mutation rates were used to estimate the age of STR variation:

-an average of 6.9×10^{-4} per 25 years across all loci determined by Zhivotovsky et al. (2004) by identifying the variation within a clade within a population that has undergone a documented recent expansion, using the 9 (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS439) of the 15 NRY STRs for which repeat sizes were obtained in this thesis.

-a per locus mutation rate (see Supplementary Table NRYSTRdate) determined using a review of the pedigree studies collected up until 2009 by the Y-chromosome Haplotype Reference Database (YHRD, http://www.yhrd.org/) (Willuweit and Roewer 2007).

This resource provides mutation rate estimates for 14 (excludes DYS388) of the 15 NRY STRs for which repeat sizes were obtained in this thesis.

Age of STR variation estimates were determined using the MS1 6 STR and the full 15 STR datasets using the Zhivotovsky et al. (2004) average mutation rate, and likewise, although excluding DYS388, using 5 of the MS1 STR dataset and 14 of the 15 full STR dataset using the per locus mutation rates provided by YHRD (see Supplementary Table NRYSTRdate).

2.3.1.4 Nucleotide diversity (π)

Nucleotide diversity (π, the probability of randomly sampling two different homologous nucleotides from a population) over L loci, as well as its variance, was estimated using the formulae of Nei (1987), as defined by $\pi = (n(\Sigma\chi_i\chi_j\pi_{ij})/(n-1))/L$, where n is the sample size, χ_I and χ_j are the frequencies for the ith and jth sequences, and π_{ij} is the proportion of different nucleotides between the ith and jth sequences. The standard deviation (s.d.) is given as the square root of the variance. Estimation of nucleotide diversity was performed using Arlequin 3.1 software (Excoffier et al. 2005).

2.3.2 Genetic distance metrics

2.3.2.1 F_{ST}

Population pairwise F_{ST} distances (Reynolds et al. 1983) of haplotype and haplogroup data was computed using Arlequin 3.1 software (Excoffier et al. 2005). F_{ST} can be simply defined as $F_{ST} = (h_T–h_S)/h_T$, where h_T is the expected total gene diversity of the metapopulation and h_S is the mean gene diversity of the sub-populations (Nei 1987). Significance of F_{ST} values was assessed by a non-parametric permutation method, which consisted of permuting haplotypes amongst populations and re-calculating F_{ST} to create a null distribution (Excoffier et al. 1992), of which 10,000 rounds of permutation were performed.

2.3.2.2 R_{ST}

Population pairwise R_{ST} distances (Goldstein et al. 1995; Slatkin 1995) of STR haplotypes was computed using Arlequin 3.1 software (Excoffier et al. 2005). R_{ST} can be defined as $R_{ST} = (S_M – S_W)/S_M$, where S_M is the total STR repeat size variance of the metapopulation over all loci and S_W is the weighted mean STR repeat size variance of

the sub-populations over all loci (Slatkin 1995). Significance of R_{ST} values was assessed using the same non-parametric permutation method used above for F_{ST}.

2.3.2.3 Kimura 2-parameter (K2P)

Population pairwise genetic distances using the Kimura 2-parameter (K2P) model of mtDNA HVS1 haplotypes was computed using Arlequin 3.1 software (Excoffier et al. 2005). The two parameters of K2P are the total frequency of transition type nucleotide substitutions P, and the total frequency of transversion type nucleotide substitutions Q, which determines the evolutionary distance per site, K, when comparing two sequences as $K = -^{1}/_{2} \ln ((1-2P-Q) \sqrt{(1-2Q)})$ (Kimura 1980). A value for the gamma distribution parameter of 0.47 was used to account for the differential rates of substitutions along the HVS1 region of the mtDNA (Wakeley 1993). Significance of K2P values was assessed using the same non-parametric permutation method used above for F_{ST}

2.3.3 Exact Tests of Population Differentiation (ETPD)

Assessment of whether the differences in the frequencies of haplotypes and haplogroups between pairs of populations was significant was determined using an Exact Test of Population Differentiation (ETPD), with 10,000 Markov steps (Raymond and Rousset 1995; Goudet et al. 1996) implemented in Arlequin 3.1 (Excoffier et al. 2005). This test is analogous to a Fisher's Exact Test (Lee et al. 2004), but with the 2x2 contingency table extended to encompass the greater number of populations and haplotypes being compared. The significance of the test is assessed by utilising a Markov chain random walk exploring the space of all possible contingency tables, with the p value as the proportion of visited tables with a probability less than or equal to the observed contingency table (Excoffier et al. 2005).

2.3.4 Principal Coordinates Analysis (PCO)

Principal coordinate analysis (PCO) (Gower 1966) was performed on pairwise genetic distance matrices using R statistical software package (www.R-project.org), implementing the 'cmdscale' function from the 'mva' package. PCO enables the visualisation of multi-dimensional data in a limited number of dimensions (such as a 2D plot), by computing the vectors that maximise the variability in the data.

2.3.5 Mantel tests

Correlations between corresponding values in different distance matrices were investigated using Mantel and Partial Mantel tests (Sokal and Rohlf 1994). Mantel and

Partial Mantel tests were performed using R statistical software package (www.R-project.org), using the Pearson product-moment method in the 'vegan' package. Significance of the correlation was assessed by permuting rows and columns of the matrices 10,000 times.

2.3.6 Network construction

Network 4.516 (www.fluxus-engineering.com) was used to construct median joining networks. Networks constructed using mtDNA HVS1 haplotypes used equal weighting for transition and transversion nucleotide changes, while networks constructed using STR haplotypes were weighted proportionally to the inverse integer of the repeat size variance for each STR loci (Chiaroni et al. 2010).

2.3.7 Ethnographic diversity and distance

In order to see to what extent diversity in language spoken, or diversity in ancestral ethnicity for samples in an ethnic group might be associated with genetic diversity and distance, ethnographic (linguistic or ethnic) distances were estimated. Ethnographic diversity was estimated in the same way as Nei's (1987) gene diversity (see section 2.3.1.1), substituting either frequency of first language, or frequency of parental or grandparental ethnicity as appropriate for allele frequency. Ethnographic distance was estimated in the same way as was Reynolds et al. (1983) F_{ST} (see section 2.3.2.1), substituting diversity of first language or diversity of ethnicity (calculated as described above) as appropriate for gene diversity, estimated as above.

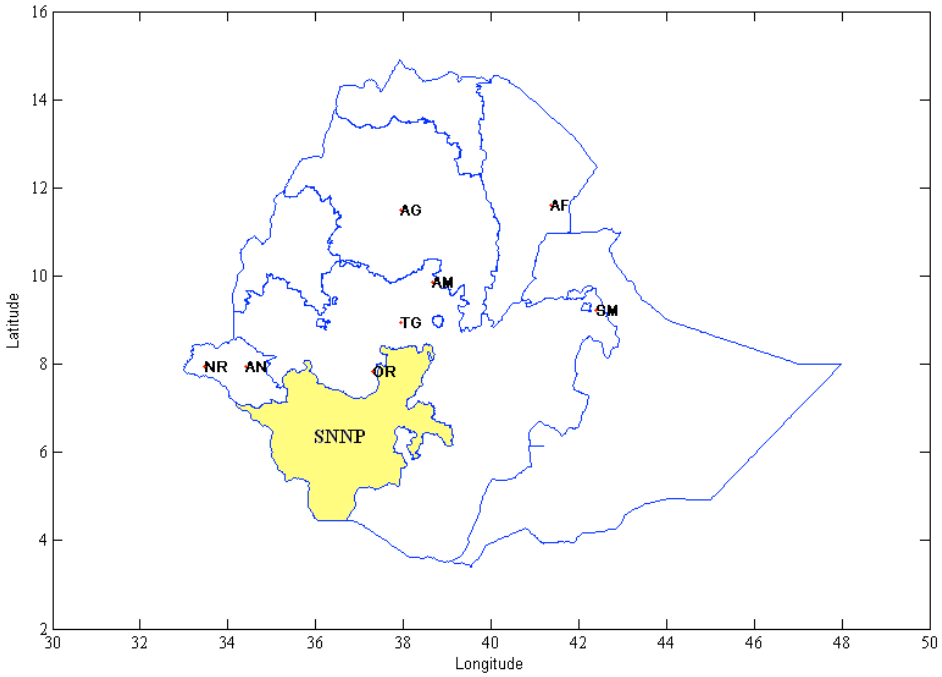
2.4 Ethnic group codes and collection locations

2.4.1 Ethnic group codes and weighted mean collection location coordinates

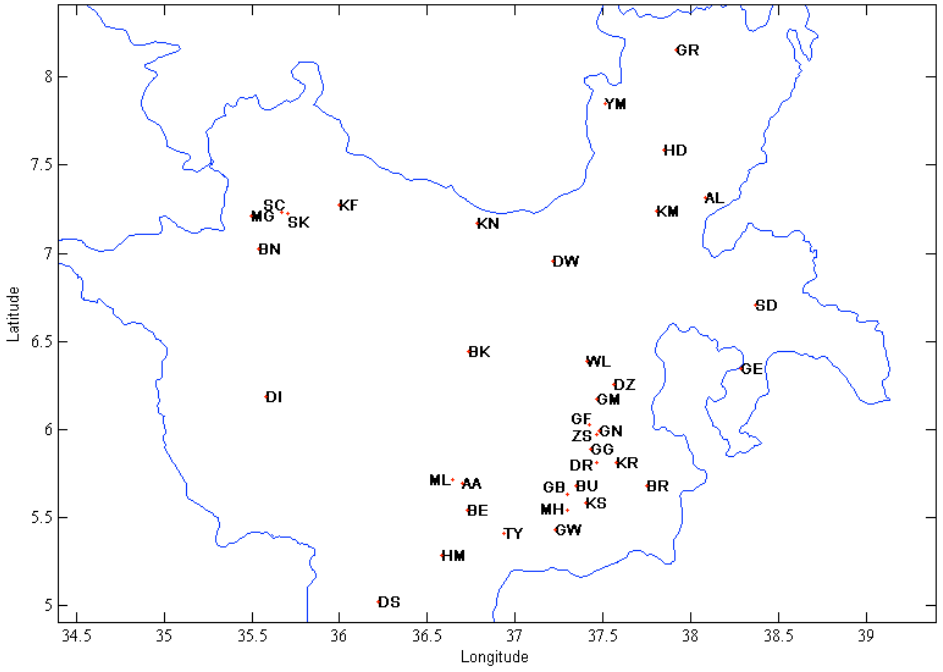
Ethnic code	Ethnic group	No. of samples	Weighted mean* latitude (° decimal)	Weighted mean* longitude (° decimal)	Weighted mean* elevation (m)
AA	Ari	117	5.69	36.70	1348
AF	Afar	112	11.60	41.36	379
AG	Agew	269	11.50	37.95	2063
AL	Alaba	110	7.32	38.08	1790
AM	Amhara	396	9.87	38.66	2088
AN	Anuak	108	7.95	34.41	500
BE	Bena	124	5.54	36.73	1528
BK	Basketo	113	6.44	36.74	1662
BN	Bench	127	7.03	35.54	1334
BR	Burji	119	5.68	37.75	1602
BU	Busa	126	5.68	37.35	2072
DI	Dizi	132	6.19	35.58	2104
DR	Dirasha	108	5.81	37.46	1679
DS	Dasanach	105	5.02	36.22	611
DW	Dawuro	117	6.95	37.22	2117
DZ	Dorze	104	6.25	37.56	2780
GB	Gobeze	113	5.63	37.30	1823
GE	Gedeo	122	6.35	38.28	1740
GF	Gofa	111	6.02	37.42	1358
GG	Ganjule	109	5.89	37.43	1175
GM	Gamo	209	6.17	37.46	1535
GN	Genta	113	5.99	37.48	1521
GR	Gurage	152	8.16	37.92	2048
GW	Gewada	117	5.43	37.23	1530
HD	Hadiya	127	7.58	37.85	2258
HM	Hamer	112	5.29	36.58	1166
KF	Kefa	120	7.28	35.99	1794
KM	Kembata	117	7.24	37.80	2053
KN	Konta	107	7.17	36.78	1663
KR	Kore	108	5.81	37.58	1381
KS	Konso	120	5.58	37.40	1150
MG	Mejenger	115	7.21	35.50	1203
MH	Mashile	130	5.54	37.29	1724
ML	Maale	119	5.71	36.64	1456
NR	Nuer	118	7.96	33.47	347
OR	Oromo	149	7.84	37.31	1758
SC	Shekecho	125	7.23	35.67	1384
SD	Sidama	126	6.71	38.37	1746
SK	Sheko	113	7.22	35.71	1335
SM	Somali	108	9.23	42.40	1543
TG	Tigray	66	8.93	37.94	1722
TY	Tsemay	114	5.41	36.94	878
WL	Wolayta	110	6.39	37.41	1737
YM	Yem	108	7.85	37.51	2499
ZS	Zayse	111	5.97	37.46	1286

*values weighted by the number of samples collected from each location for an ethnic group.

2.4.2 Map of weighted mean collection locations for ethnic groups, excluding those in the SNNP province



2.4.3 Map of weighted mean collection locations for ethnic groups in SNNP province



Chapter 3: Variation in NRY and mtDNA in Ethiopia

3.1 How much diversity is there – within populations and between populations?

3.1.1 NRY diversity

The 5756 Ethiopian samples successfully typed comprised 526 NRY haplotypes at the UEP + MS level distributed over eight NRY haplogroups (see Supplementary Table NRY), of which 232 (44.1%) were singletons (Figure 3.1). Using the nomenclature proposed by Karafet et al. (2008), the modal haplogroup was E*(xE1b1a), accounting for 69.2% of the data. It was at highest frequency in the Ganjule (GG, 93.6%), and lowest in the Anuak (AN, 16.7%). The second most frequent haplogroup was J, accounting for 12.0% of the data, and present at highest frequency in the Shekecho (SC, 52.0%). It was not seen in the Gewada (GW), Mejenger (MG), Mashile (MH), Tsemay (TY) and Zayse (ZS) datasets.

Figure 3.1 Counts of NRY UEP-MS haplotypes and frequency of occurrence in dataset

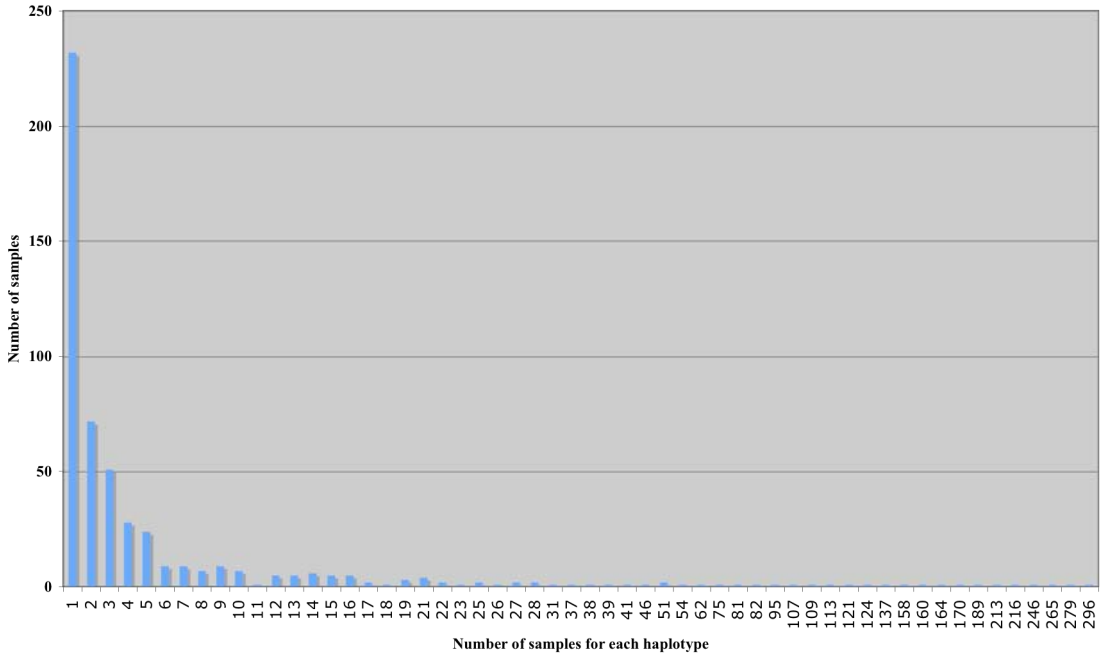
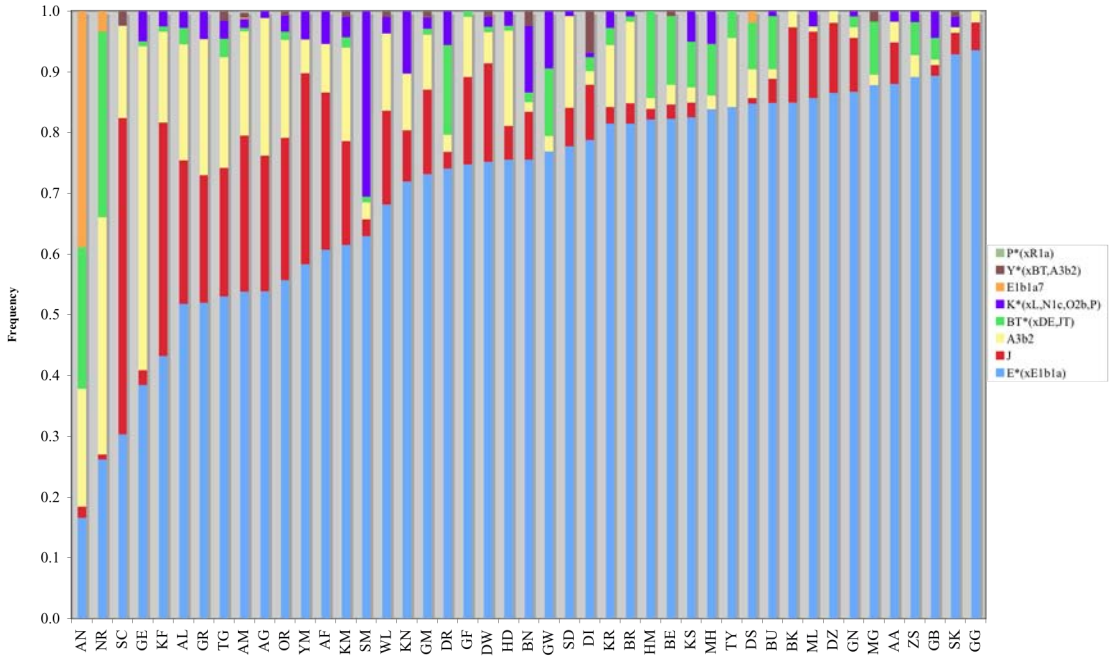


Figure 3.2 Frequencies of NRY haplogroups found in Ethiopian ethnic groups (ordered by increasing NRY haplogroup gene diversity)



Gene diversity (*h*, Table 3.1), at haplogroup level ranged from 0.123 (Ganjule, GG) to 0.736 in the Anuak (AN). The mean *h* value across all ethnic groups was 0.414, and the median value 0.403. The two groups with lowest diversity (Ganjule and Sheko, GG and SK respectively) are notable in having the highest frequencies of haplogroup E*(xE1b1a) (94% and 93% respectively), while the two groups with greatest diversity (Anuak and Nuer, AN and NR respectively) both have the highest frequencies of haplogroup BT*(xDE,JT) (23% and 31% respectively) as well as the highest frequencies of haplogroup E1b1a7 (39% and 3% respectively), a haplogroup almost completely absent from all other ethnic group datasets.

Gene diversity based on UEP-MS haplotypes ranged from its lowest at 0.743 (Hadiya, HD) to its highest at 0.972 (Wolayta, WL). Gene diversity in ethnic groups at the haplotype level was comparable to the levels of diversity recently observed in the linguistically diverse Cross River region of Nigeria (Veeramah et al. 2010), where gene diversity values ranged from 0.911 to 0.973 in the 24 clans surveyed. In the Ethiopian ethnic groups, a mean *h* value across all ethnic groups of 0.920 was observed, and 36 of the 45 ethnic groups had *h* values greater than 0.900. Mean microsatellite repeat length variance (MSV) ranged from 0.345 (Majenger, MG) to 1.315 (Tigray, TG). The mean MSV across all ethnic groups was 0.786, and the median value 0.744.

Table 3.1 Summary of the genetic diversity found within Ethiopian ethnic groups

Ethnic group code	NRY haplogroup *h*	s.d ±	NRY haplotype *h*	s.d ±	NRY MS MSV	mtDNA haplotype *h*	s.d ±	mtDNA haplotype π	s.d ±
AA	0.221	0.038	0.839	0.034	0.526	0.986	0.011	0.0241	0.0124
AF	0.560	0.047	0.940	0.022	0.851	0.987	0.011	0.0246	0.0126
AG	0.610	0.030	0.954	0.013	1.212	0.988	0.007	0.0235	0.0120
AL	0.644	0.046	0.968	0.017	1.122	0.993	0.008	0.0257	0.0131
AM	0.616	0.024	0.971	0.008	1.250	0.990	0.005	0.0233	0.0119
AN	0.736	0.042	0.912	0.027	0.851	0.992	0.009	0.0266	0.0136
BE	0.311	0.042	0.907	0.026	0.483	0.984	0.011	0.0237	0.0122
BK	0.265	0.041	0.783	0.039	0.586	0.987	0.011	0.0222	0.0115
BN	0.412	0.044	0.922	0.024	0.769	0.984	0.011	0.0221	0.0114
BR	0.319	0.043	0.934	0.023	0.861	0.986	0.011	0.0249	0.0127
BU	0.271	0.040	0.919	0.024	0.466	0.979	0.013	0.0228	0.0117
DI	0.368	0.042	0.929	0.022	0.600	0.974	0.014	0.0226	0.0116
DR	0.429	0.048	0.966	0.017	0.715	0.992	0.009	0.0251	0.0128
DS	0.276	0.044	0.939	0.023	0.744	0.992	0.009	0.0270	0.0138
DW	0.408	0.045	0.950	0.020	0.686	0.993	0.008	0.0237	0.0122
DZ	0.240	0.042	0.916	0.027	0.603	0.985	0.012	0.0262	0.0134
GB	0.199	0.038	0.937	0.023	0.606	0.975	0.015	0.0225	0.0116
GE	0.569	0.045	0.920	0.024	0.937	0.985	0.011	0.0246	0.0126
GF	0.414	0.047	0.970	0.016	0.940	0.992	0.008	0.0236	0.0121
GG	0.123	0.031	0.811	0.038	0.438	0.969	0.017	0.0257	0.0131
GM	0.438	0.034	0.955	0.014	0.905	0.994	0.005	0.0243	0.0124
GN	0.241	0.040	0.919	0.026	0.644	0.991	0.009	0.0261	0.0133
GR	0.638	0.039	0.964	0.015	1.167	0.989	0.008	0.0228	0.0117
GW	0.390	0.045	0.955	0.019	0.443	0.970	0.016	0.0241	0.0123
HD	0.403	0.044	0.743	0.039	0.587	0.992	0.008	0.0252	0.0129
HM	0.307	0.044	0.931	0.024	0.502	0.980	0.013	0.0239	0.0123
KF	0.647	0.044	0.967	0.016	1.279	0.972	0.015	0.0218	0.0112
KM	0.572	0.046	0.956	0.019	1.030	0.995	0.007	0.0259	0.0132
KN	0.460	0.048	0.954	0.020	0.627	0.993	0.008	0.0234	0.0120
KR	0.326	0.045	0.853	0.034	0.598	0.990	0.009	0.0251	0.0128
KS	0.313	0.042	0.946	0.021	0.450	0.987	0.010	0.0261	0.0133
MG	0.222	0.039	0.745	0.041	0.345	0.962	0.018	0.0233	0.0120
MH	0.289	0.040	0.946	0.020	0.744	0.975	0.014	0.0236	0.0121
ML	0.255	0.040	0.937	0.022	0.663	0.984	0.011	0.0240	0.0123
NR	0.691	0.043	0.940	0.022	0.838	0.990	0.009	0.0264	0.0135
OR	0.612	0.040	0.959	0.016	1.308	0.995	0.006	0.0240	0.0123
SC	0.618	0.043	0.947	0.020	1.312	0.980	0.013	0.0215	0.0111
SD	0.371	0.043	0.957	0.018	0.923	0.993	0.007	0.0254	0.0130
SK	0.136	0.032	0.882	0.030	0.382	0.982	0.012	0.0246	0.0126
SM	0.513	0.048	0.891	0.030	1.001	0.983	0.012	0.0230	0.0118
TG	0.648	0.059	0.969	0.021	1.315	0.991	0.012	0.0234	0.0121
TY	0.278	0.042	0.886	0.030	0.502	0.984	0.012	0.0239	0.0123
WL	0.499	0.048	0.972	0.016	1.067	0.996	0.006	0.0248	0.0127
YM	0.561	0.048	0.911	0.027	0.988	0.979	0.014	0.0248	0.0127
ZS	0.202	0.038	0.912	0.027	0.524	0.962	0.018	0.0243	0.0124

Figure 3.3 NRY UEP haplogroup gene diversity values in Ethiopian ethnic groups (ordered by increasing gene diversity)

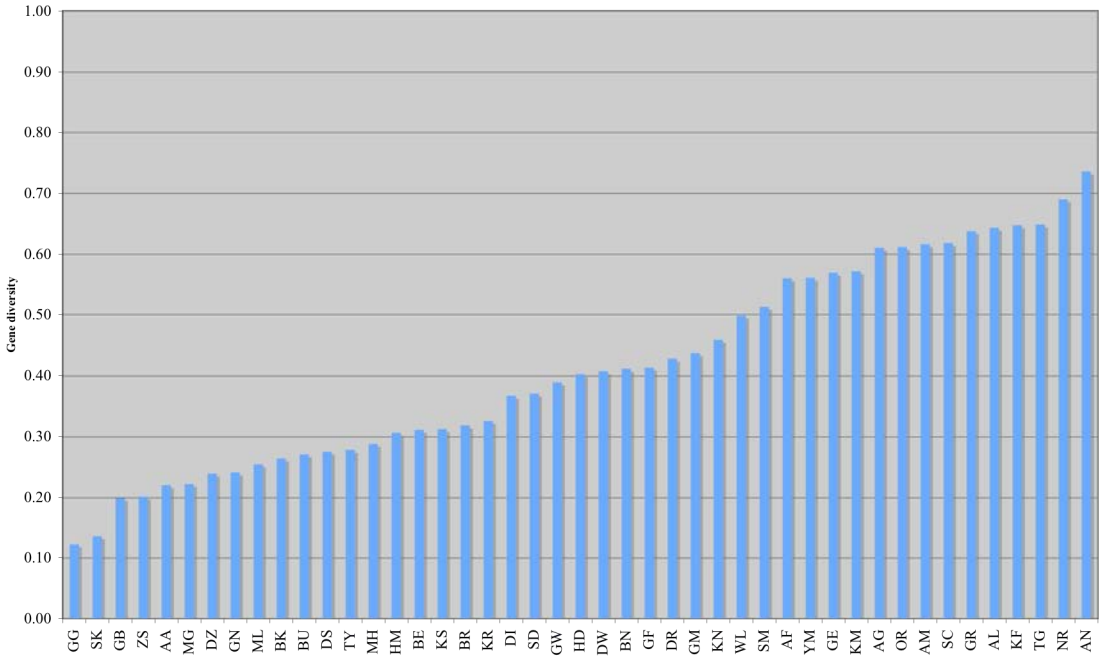


Figure 3.4 NRY UEP-MS haplotype gene diversity values in Ethiopian ethnic groups (ordered by increasing gene diversity)

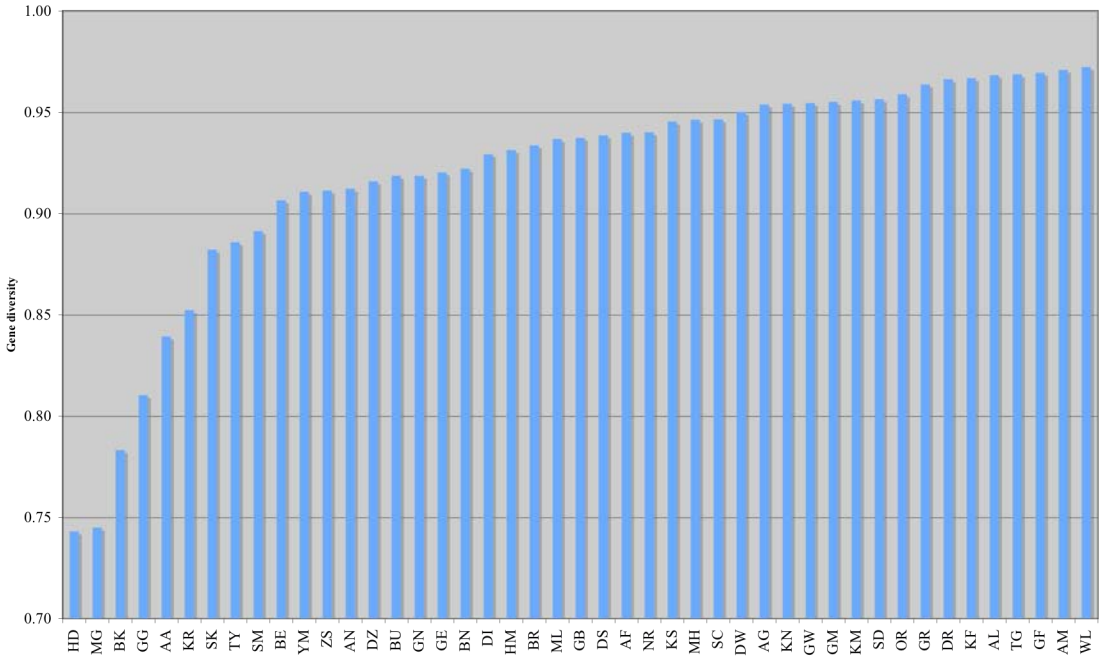
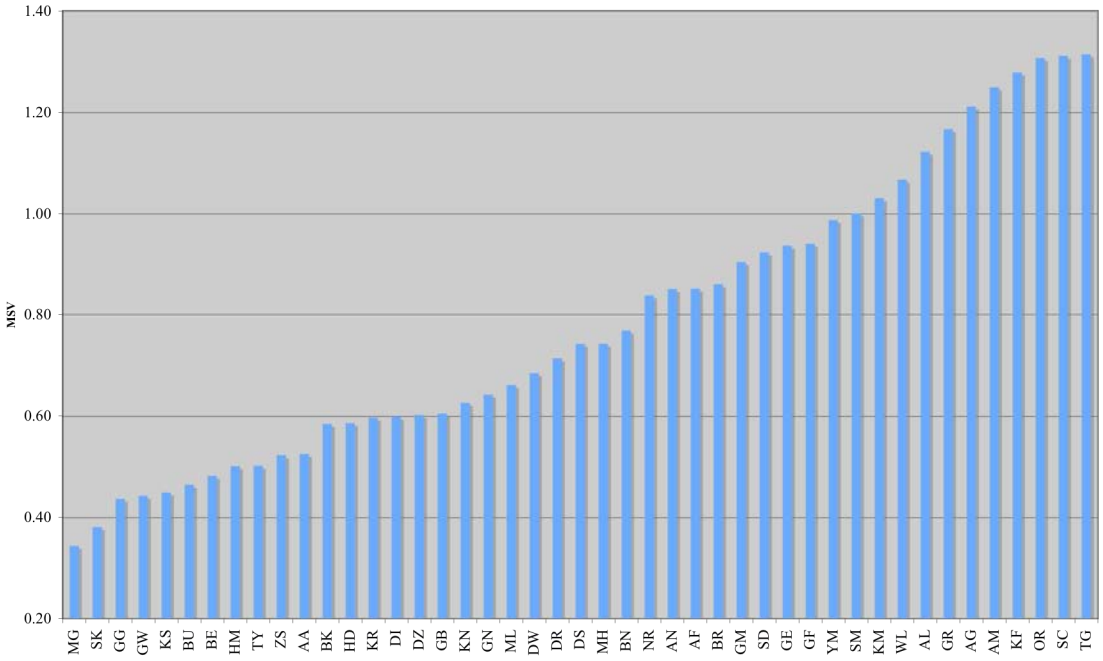


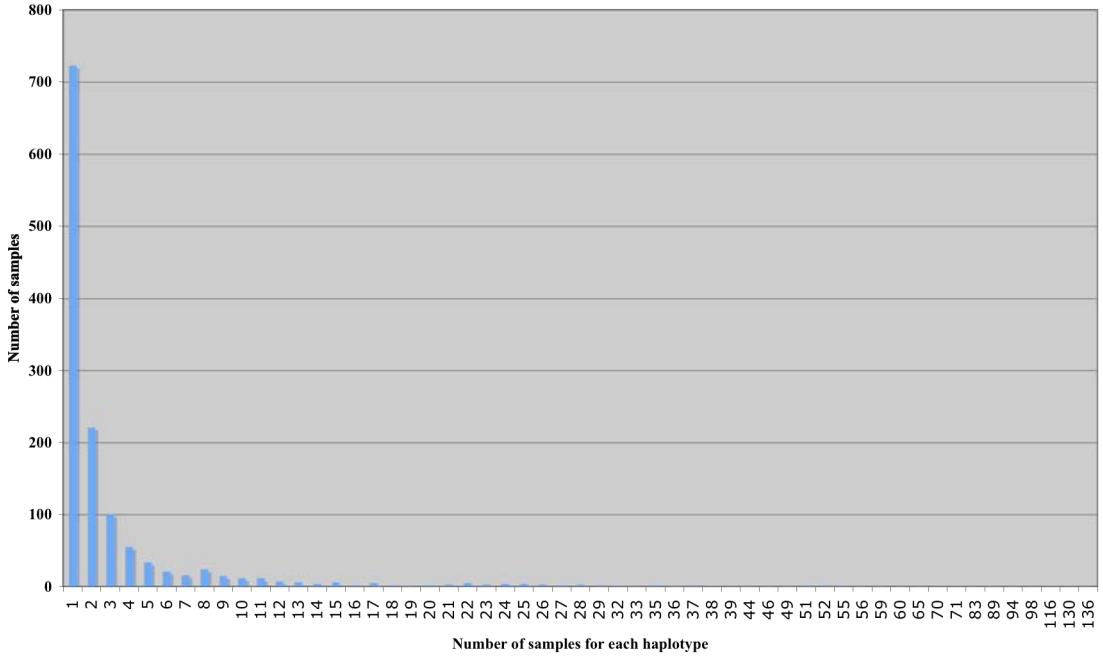
Figure 3.5 NRY mean microsatellite repeat length variance in Ethiopian ethnic groups (ordered by increasing variance)



3.1.2 mtDNA diversity

A total of 5756 successfully sequenced samples were distributed among 1328 mtDNA HVS1 defined haplotypes (see Supplementary Table mtDNA), of which 723 (54.4%) were singletons (Figure 3.6).

Figure 3.6 Counts of mtDNA HVS1 haplotypes and frequency of occurrence in dataset



Gene diversity based on the mtDNA HVS1 haplotype were comparable to those estimated for the Cross River region of Nigeria, where they ranged from 0.978 to 1.000, with a mean *h* of 0.991 in the 24 clans (Veeramah et al. 2010). In the 45 Ethiopian ethnic groups, mtDNA haplotype *h* values ranged from 0.962 found in both the Majenger and the Zayse (ZS) to 0.996 (Wolayta, WL), with a mean *h* value across all ethnic groups of 0.985, and the median value 0.987. Nucleotide diversity (π) values ranged from 0.0215 (Shekecho, SC) to 0.0270 (Dasanach, DS). The mean and the median nucleotide diversity across all ethnic groups were both 0.0240.

Figure 3.7 mtDNA HVS1 haplotype gene diversity values in Ethiopian ethnic groups (ordered by increasing gene diversity)

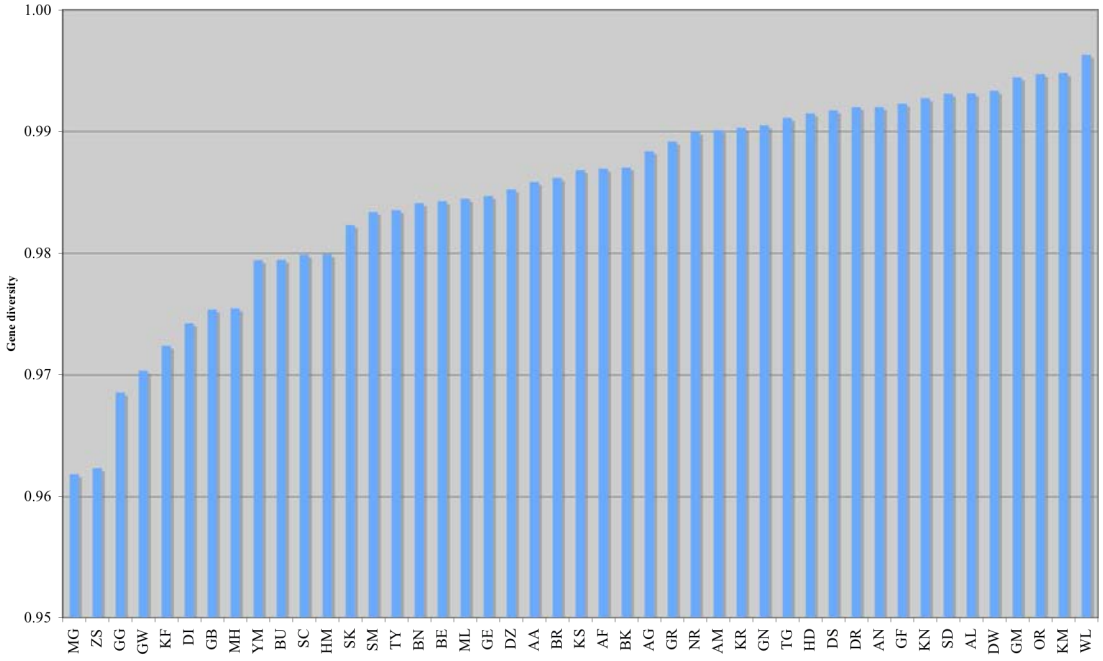
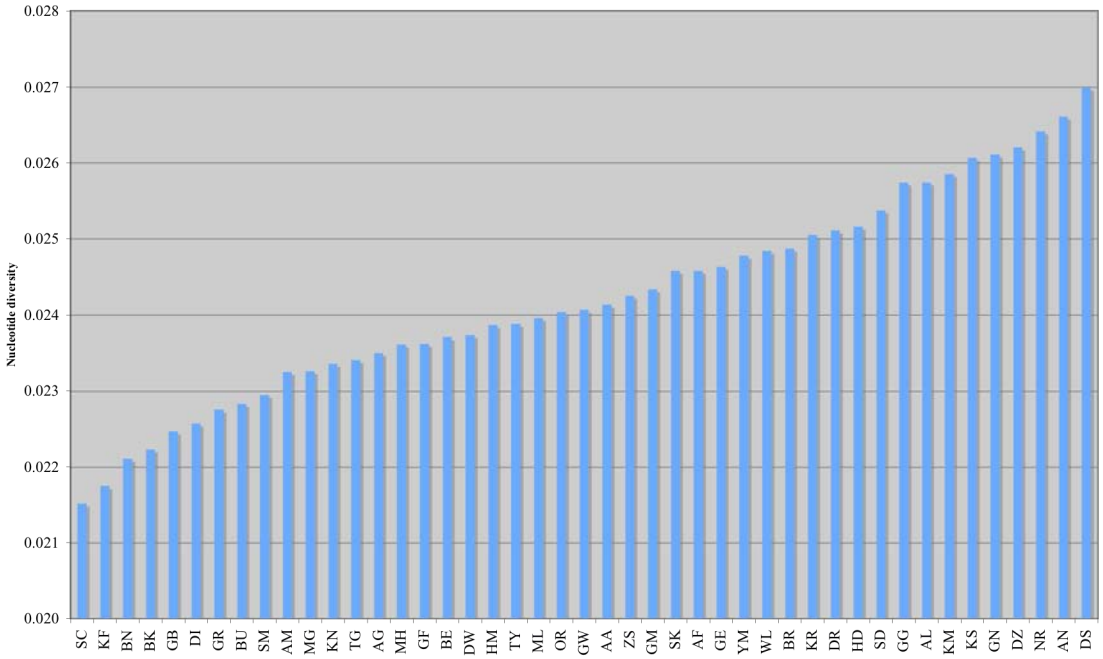


Figure 3.8 mtDNA HVS1 haplotype nucleotide diversity values in Ethiopian ethnic groups (ordered by increasing nucleotide diversity)



3.1.3 Correlation between NRY and mtDNA diversity

Significant positive correlation of ranked values of NRY UEP-MS and mtDNA HVS1 haplotypes in each ethnic group were observed (Spearman's *r* of 0.5826, $p<0.0001$). For a given ethnic group, there was approximately twice the number of samples per NRY haplotypes compared to the number of samples per mtDNA haplotypes, with mean values across all ethnic groups of 3.203 and 1.686 for NRY and mtDNA samples per haplotype per dataset respectively (Figure 3.9). There was also a significant (each $p<0.01$) positive rank correlation between NRY UEP-MS haplotype gene diversity and mtDNA HVS1 haplotype gene diversity (Figure 3.10), as well as almost all other measures of NRY and mtDNA diversity (Table 3.2), with the exception of mtDNA HVS1 nucleotide diversity which was observed to be only significantly correlated with mtDNA HVS1 gene diversity.

Figure 3.9 Plot of the average number of samples per haplotypes per ethnic group for NRY and mtDNA

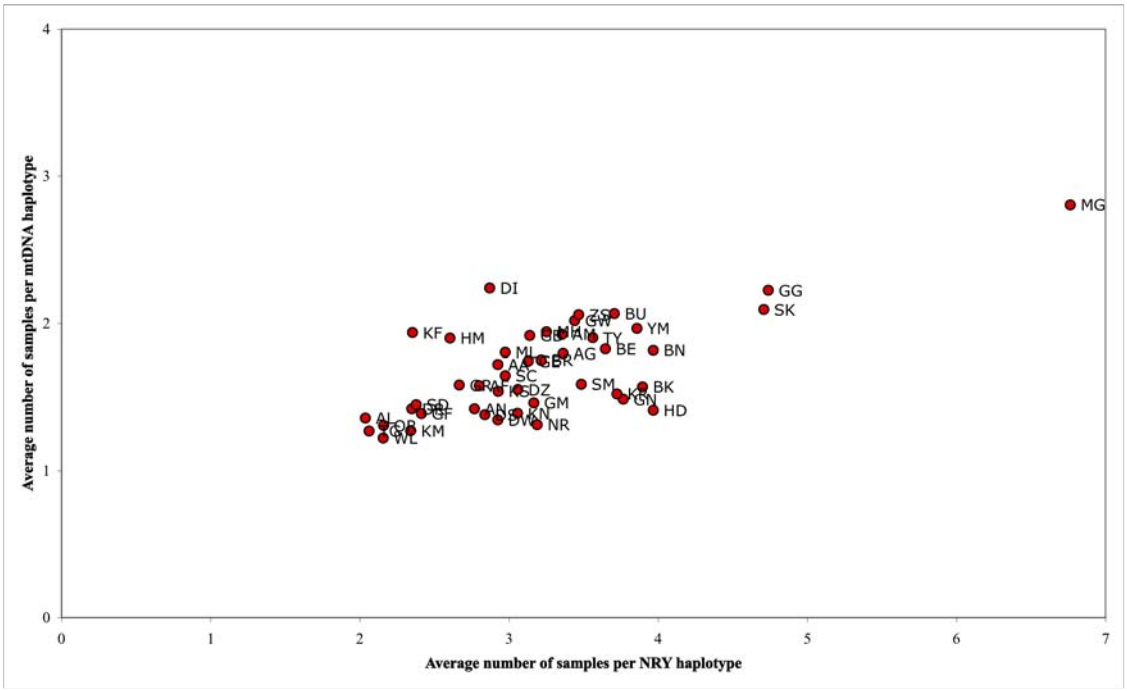


Figure 3.10 Plot of gene diversity values of NRY UEP-MS haplotypes against gene diversity of mtDNA HVS1 haplotypes for ethnic groups

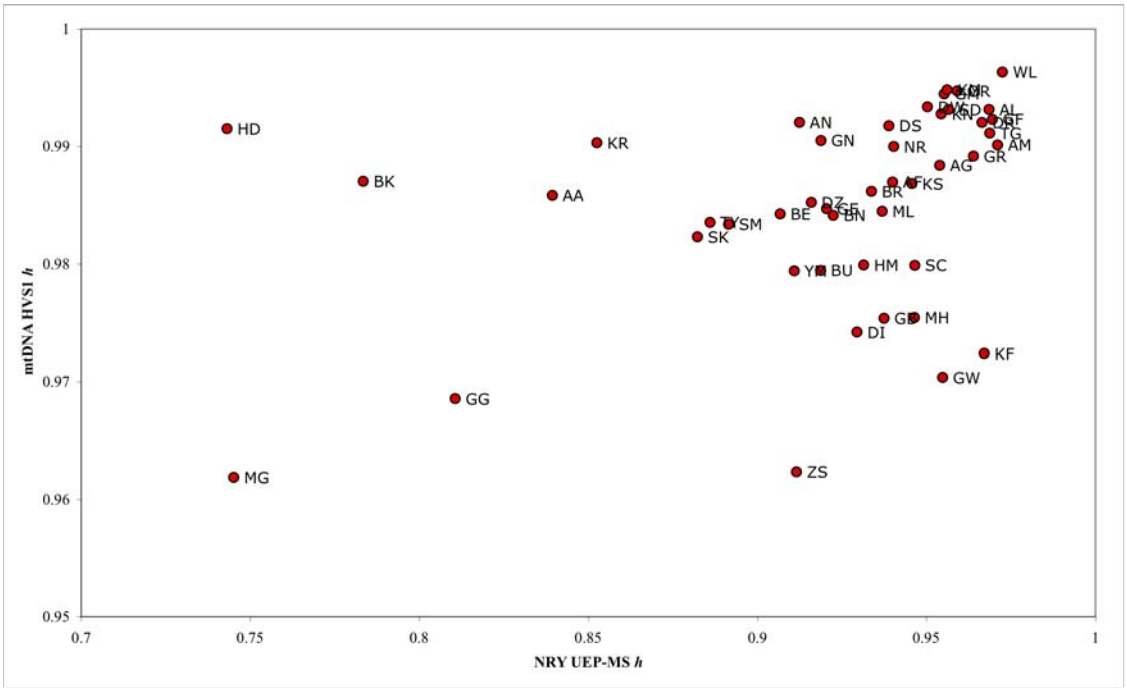


Table 3.2 Correlations between ranked diversity values in ethnic groups (p-values in upper diagonal, correlation coefficients in lower diagonal)

	NRY UEP *h*	NRY UEP-MS *h*	NRY MS MSV	mtDNA HVS1 *h*	mtDNA HVS1 π
NRY UEP *h*	*	**<0.0001**	**<0.0001**	**0.0019**	0.7070
NRY UEP-MS	0.6017	*	**<0.0001**	**0.0006**	0.5640
NRY MS MSV	0.8100	0.6742	*	**0.0017**	0.4122
mtDNA HVS1	0.4506	0.4934	0.4558	*	**0.0064**
mtDNA HVS1	-0.0576	-0.0883	-0.1253	0.4005	*

3.1.4 Exact Tests of Population Differentiation between ethnic groups

Using NRY UEP haplogroup data, Exact Tests of Population Differentiation (ETPD, Supplementary Table ETPD) showed the Anuak (AN), Gedeo (GE), Nuer (NR) and Somali (SM) to be the most distinct ethnic groups in Ethiopia, with all of the maximum 44 possible comparisons significant ($p<0.01$). The least differentiated ethnic group based on UEP haplogroup data was the Gamo (GM), with 28 out of the 44 maximum possible comparisons significant ($p<0.01$). The mean value for the number of significant comparisons across all ethnic groups was 35.47, and the median value was 35. When ETPD was performed using NRY UEP-MS haplotype data most ethnic groups were distinct from all others (Figure 3.12), with a mean value for the number of significant comparisons of 43.42, and a median value of 44. The Tigrayans (TG) were the least differentiated ethnic group using this data, with 38 of the 44 possible comparisons significant, with non-significance recorded between the Tigrayans and the Agew (AG), Alaba (AL), Amhara (AM), Gurage (GR), Kembata (KM) and Oromo (OR) (Figure 3.13).

Figure 3.11 Plot showing total number of significant ETPD comparisons for Ethiopian ethnic groups using NRY UEP haplogroup frequencies (ordered by increasing value)

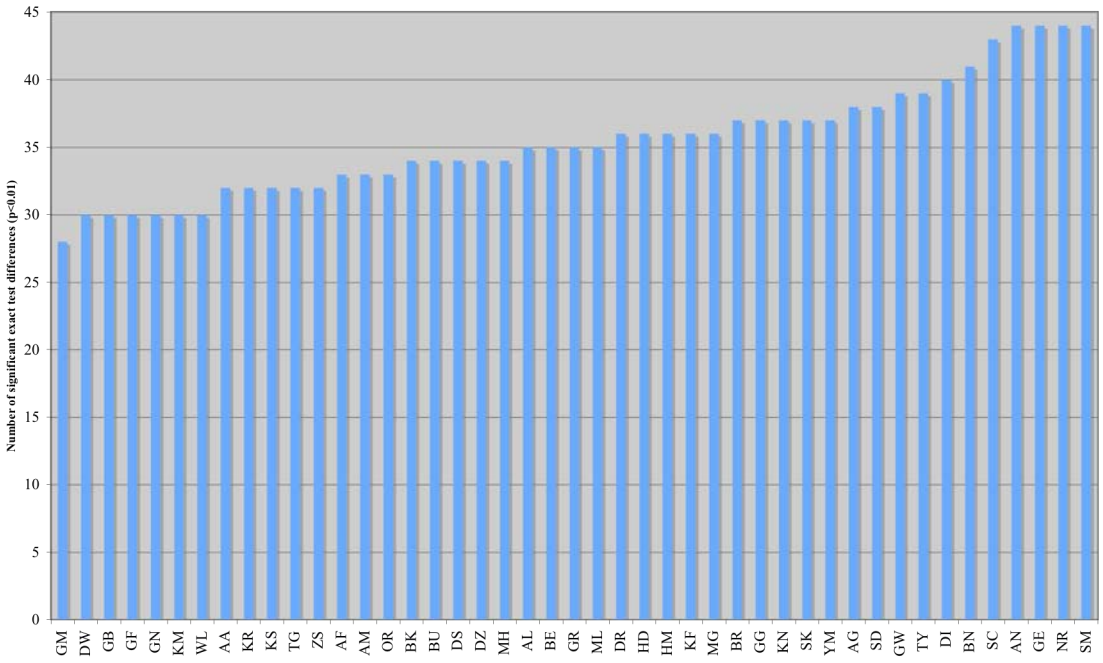


Figure 3.12 Plot showing total number of significant ETPD comparisons for Ethiopian ethnic groups using NRY UEP-MS haplotype frequencies (ordered by increasing number)

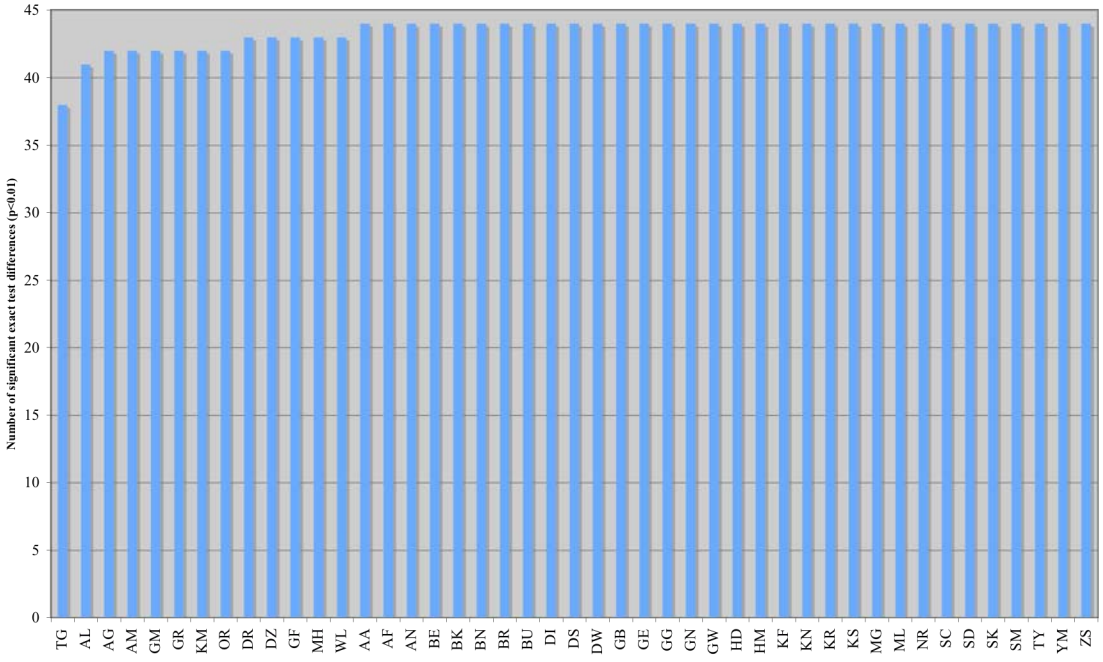
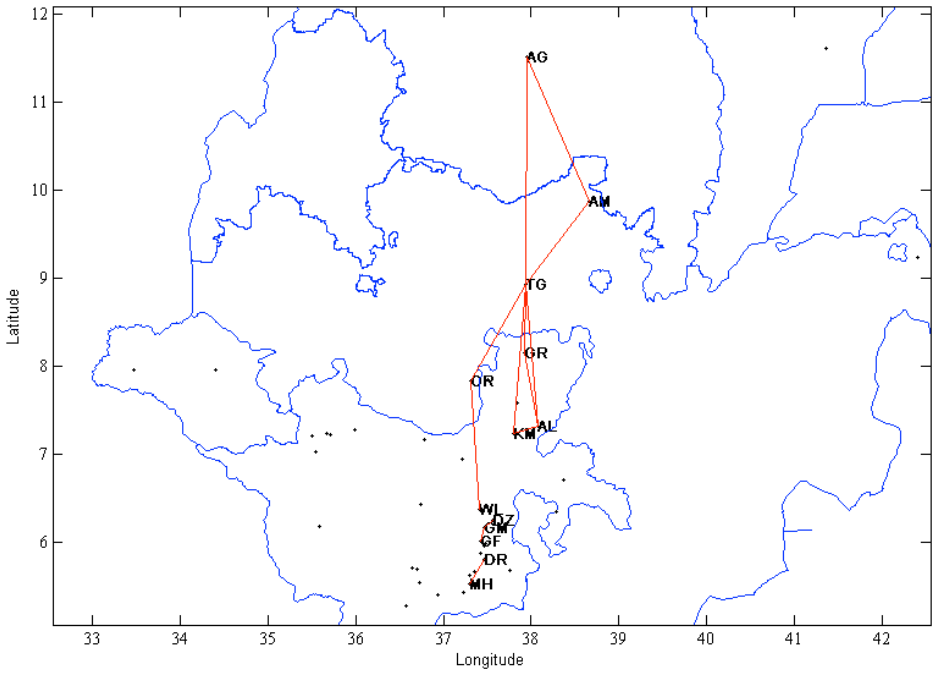


Figure 3.13 Maps showing non-significant (p>0.01) ETPD comparisons (indicated by red lines) between Ethiopian ethnic groups (indicated by dots) using NRY UEP-MS haplotype frequencies



ETPD performed using mtDNA HVS1 haplotype data generally gave high values for the number of significant comparisons, with a mean value for the number of significant comparisons of 41.87, and a median value of 43. The Wolayta (WL) were the least differentiated ethnic group using this data (Figure 3.14), with 32 of the possible 44 comparisons significant, with non-significance shown between the Wolayta and the Alaba (AL), Basketo (BK), Dawuro (DW), Gofa (GF), Gamo (GM), Genta (GN), Hadiya (HD), Kembata (KM), Konta (KN), Kore (KR), Oromo (OR) and the Tigray (TG) (Figure 3.15).

Figure 3.14 Plot showing total number of significant ETPD comparisons for Ethiopian ethnic groups using mtDNA haplotype frequencies (ordered by increasing number)

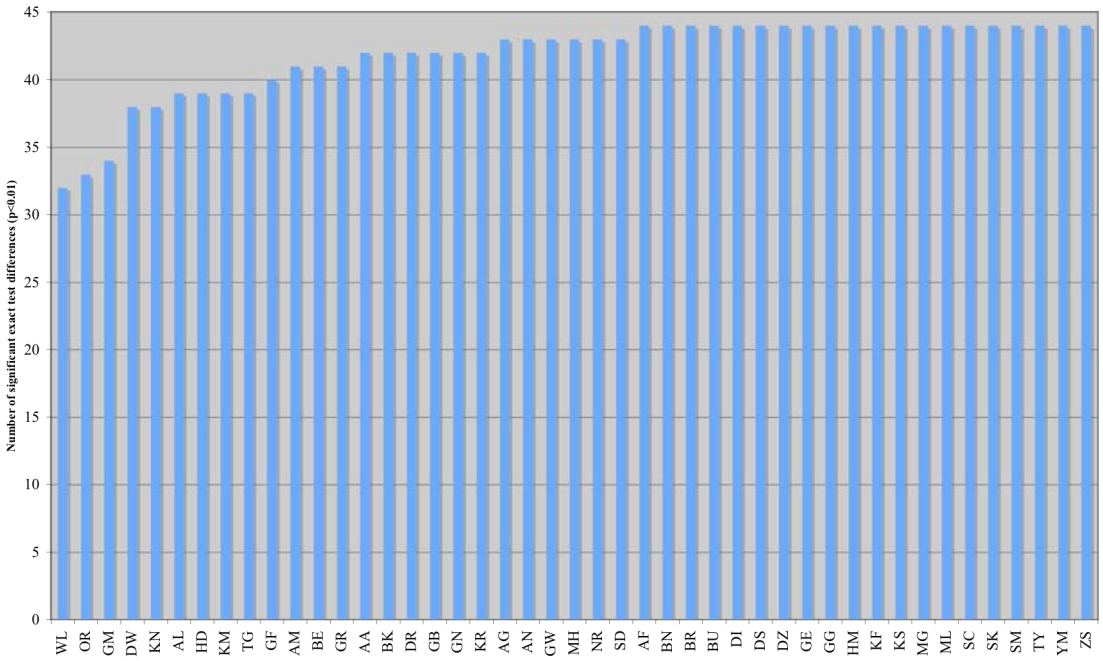
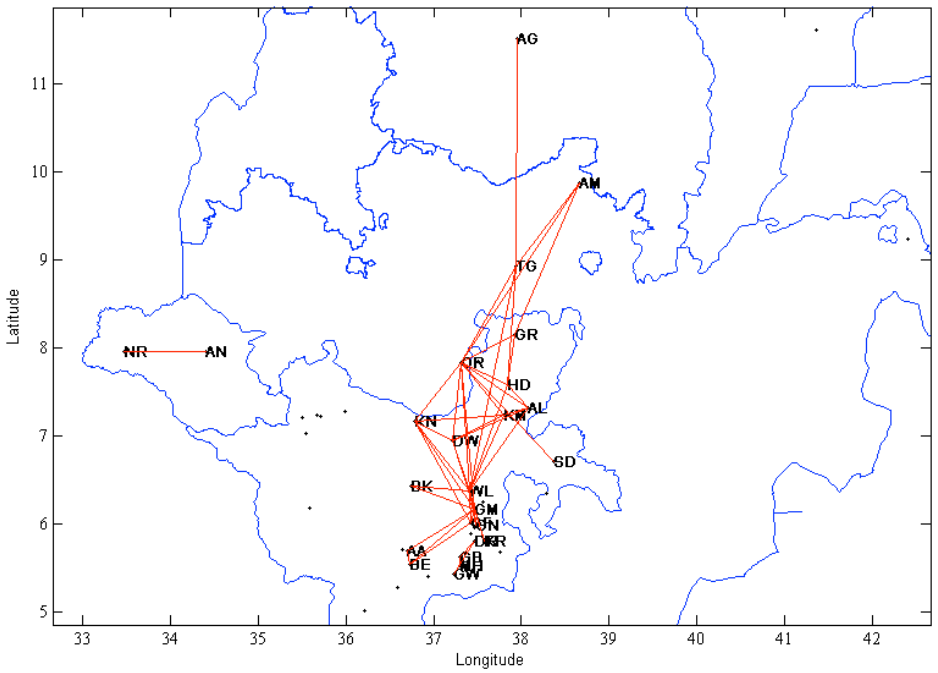


Figure 3.15 Maps showing non-significant (p>0.01) ETPD comparisons (indicated by red lines) between Ethiopian ethnic groups (indicated by dots) using mtDNA HVS1 haplotype frequencies



3.1.5 Genetic distances between ethnic groups

The first two principal coordinates determined for the pairwise genetic distances between Ethiopian ethnic groups are shown in the PCO plots below. The complete table of all pairwise genetic distances can be found in Supplementary Table Distances.

The PCO plot of pairwise Fst using NRY UEP haplogroups (Figure 3.16) show the majority of the Ethiopian ethnic groups loosely clustering together, with the Anuak (AN), Nuer, (NR) Gedeo (GE) and Shekecho (SC) appearing as outliers in the plot. These four ethnic groups have the highest frequencies of some of the less common haplogroups found in Ethiopia, namely haplogroups E1b1a7, BT*(xDE,JT), A3b2 and J found at highest frequency in the Anuak, Nuer, Gedeo and Shekecho respectively. After removal of these groups (Figure 3.17), the Kefa (KF) appear as outliers at the lower extremity of PCO1, and the Somali (SM) as outliers at the lower extremity of PCO2. The Kefa have the second highest frequency of haplogroup J, after the Shekecho, as well as one of the lowest frequencies of haplogroup E*(xE1b1a), only greater in frequency than the four excluded groups. The Somali have by far the highest frequency of haplogroup K*(xL,N1c,O2b,P) at 38.9%, with no other ethnic group having a frequency of this haplogroup greater than 3.4%. Ignoring the Kefa and Somali, the distribution of ethnic groups along PCO1 appears to reflect the frequency of haplogroup E*(xE1b1a), the most common haplogroup in Ethiopia, appearing at highest frequencies in the Ganjule and Sheko (GG and SK respectively, at the higher extremity of PCO1), and lowest frequencies in the Alaba, Gurage and Tigray (AL, GR, and TG respectively, towards the lower extremity of PCO1).

Figure 3.16 PCO of pairwise Fst distances between Ethiopian ethnic groups using UEP haplogroup frequencies

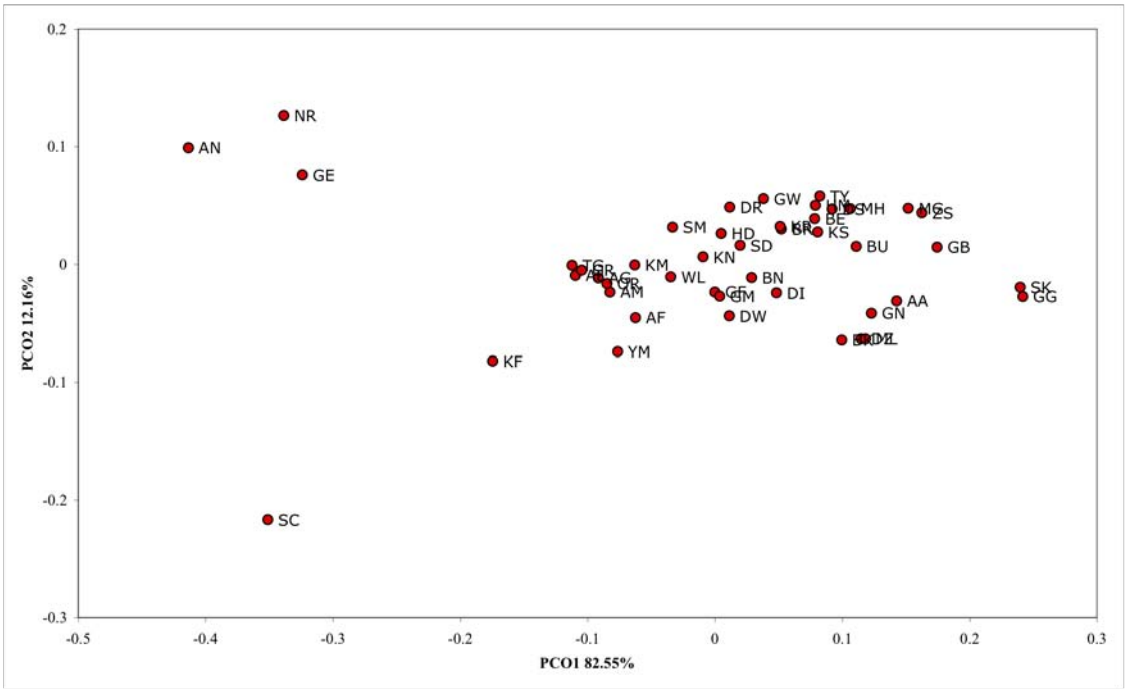
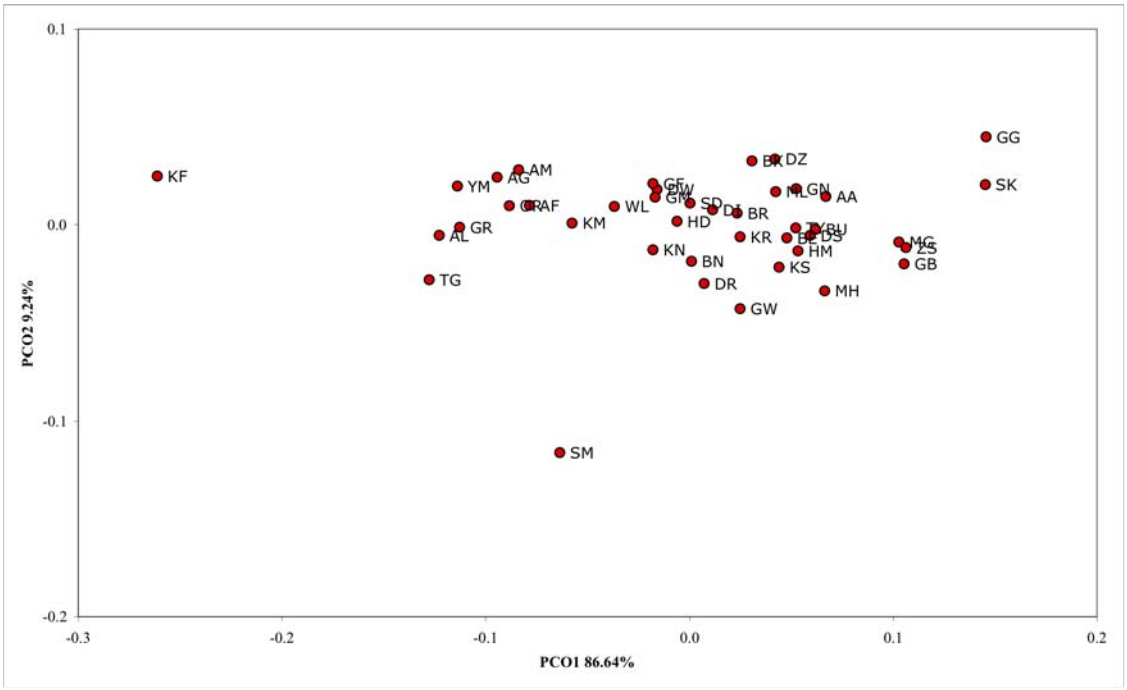


Figure 3.17 PCO of pairwise Fst distances between Ethiopian ethnic groups using UEP haplogroup frequencies, excluding the Anuak, Nuer, Gedeo and Shekecho



The PCO plot of pairwise Fst using NRY UEP-MS haplotypes (Figure 3.18) shows a loose cluster containing the majority of the ethnic groups with the Majenger (MG) and Hadiya (HD) as outliers. These two ethnic groups have by far the lowest gene diversity values out of all the ethnic groups, when estimated using UEP-MS haplotype frequencies (h=0.743 for the Hadiya, h=0.745 for the Majenger). After these groups are removed (Figure 3.19), the remaining ethnic groups spread along both PCO1 and PCO2, with a dense central cluster. The Basketo (BK) and Ari (AA) appear as outliers at the higher extremity of PCO1, whilst the Ganjule (GG), Kore (KR) and Sheko (SK) appear at the lower extremity of PCO1, and the Somali (SM), Tsemay (TY) and Anuak (AN) appear towards the lower extremity of PCO2. The nine ethnic groups mentioned above have the lowest gene diversity values amongst all the ethnic groups (h<0.900), which may explain their positioning on the PCO plot.

Figure 3.18 PCO of pairwise Fst distances between Ethiopian ethnic groups using UEP-MS haplotypes

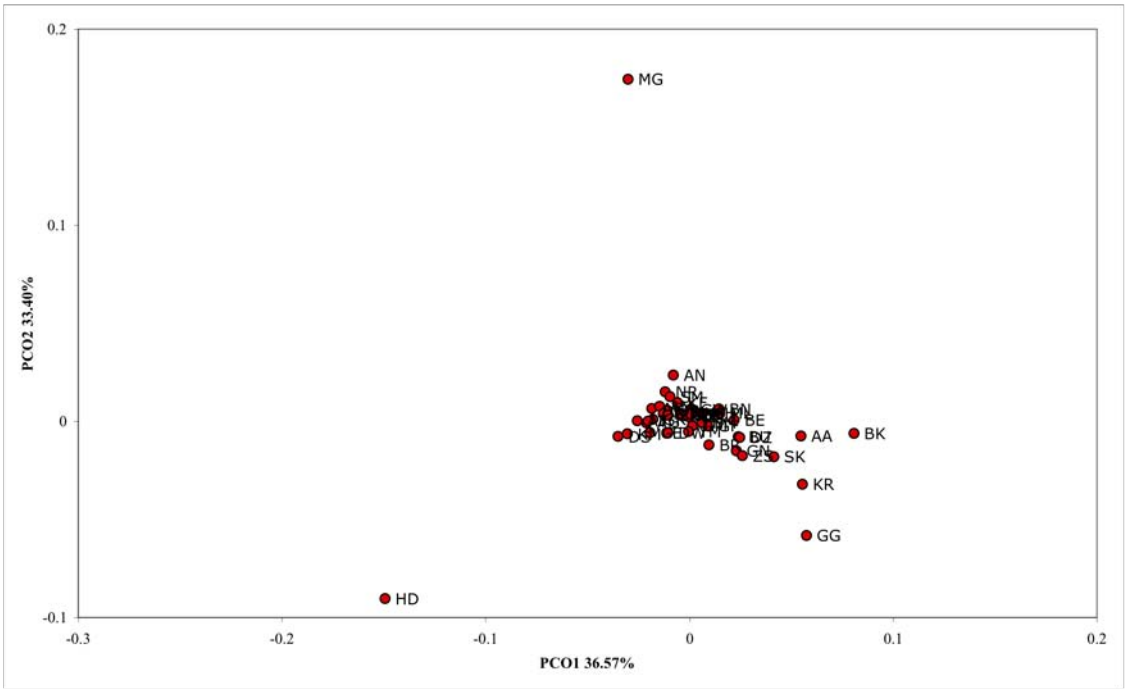
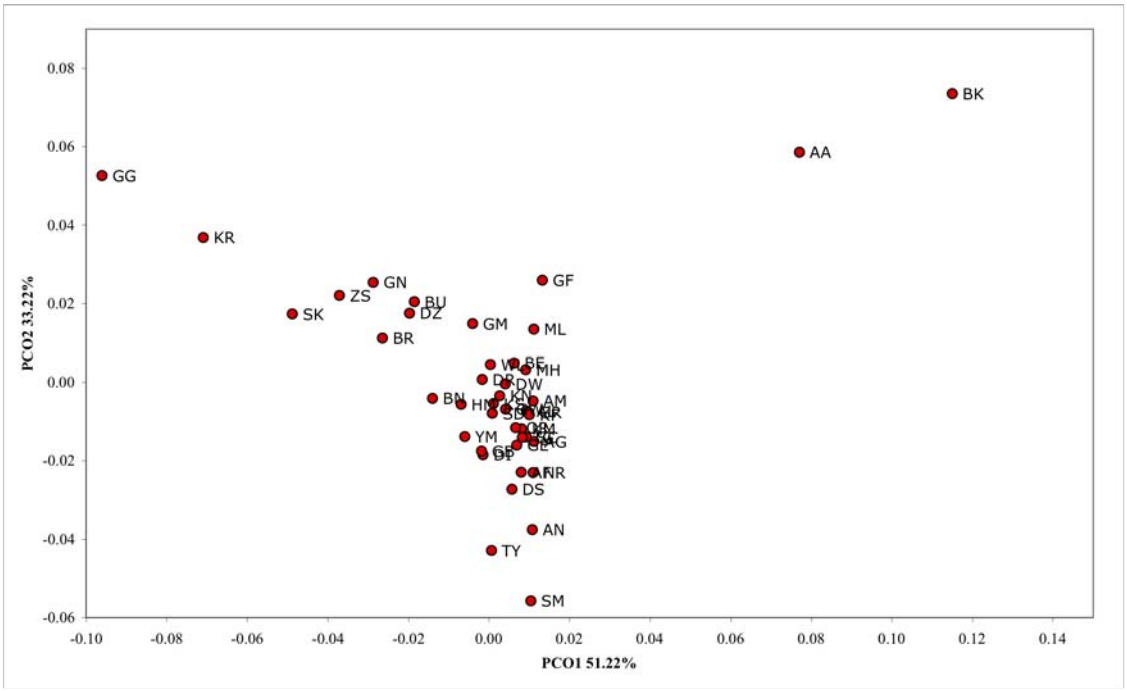


Figure 3.19 PCO of pairwise Fst distances between Ethiopian ethnic groups using UEP-MS haplotypes, excluding the Majenger and Hadiya



The PCO plot of pairwise Rst distances using MS haplotypes (Figure 3.20) shows a more diffuse spread of ethnic groups. The Gedeo (GE), Anuak (AN), Nuer (NR) and Majenger (MG) appear as outliers along PCO1, and the Somali (SM) as outliers along PCO2. The Gedeo, Anuak, Nuer and Somali have the highest frequency of the less common Ethiopian haplogroups A3b2, E1b1a7, BT*(xDE,JT) and K*(xL,N1c,O2b,P) respectively, and due to the low level of microsatellite haplotypes shared among haplogroups (4.2% of haplotypes), would be expected to generate substantial Rst distances. The Majenger on the other hand, have very low levels of diversity as measured by UEP-MS haplotype level *h* and MSV. Interestingly, 47.0% of Majenger belong to a single UEP-MS haplotype (E*(xE1b1a) 15 11 23 11 11 13) not observed in any other Ethiopian ethnic group. After removal of these five outlier ethnic groups (Figure 3.21), the remaining ethnic groups are spread more diffusely along both PCO1 and PCO2. There is however, a tight clustering in the centre of the plot which contains all of the Highland and most of the more northern SNNP ethnic groups (see section 2.4.2 and 2.4.3), and includes the Tigray (TG), Oromo (OR), Amhara (AM), Alaba (AL), Agew (AG), Kembata (KM) and Gurage (GR), with the Afar (AF), Dawuro (DW), Gofa (GF) and Wolayta (WL) ethnic groups clustering nearby.

Figure 3.20 PCO of pairwise Rst distances between Ethiopian ethnic groups using NRY MS

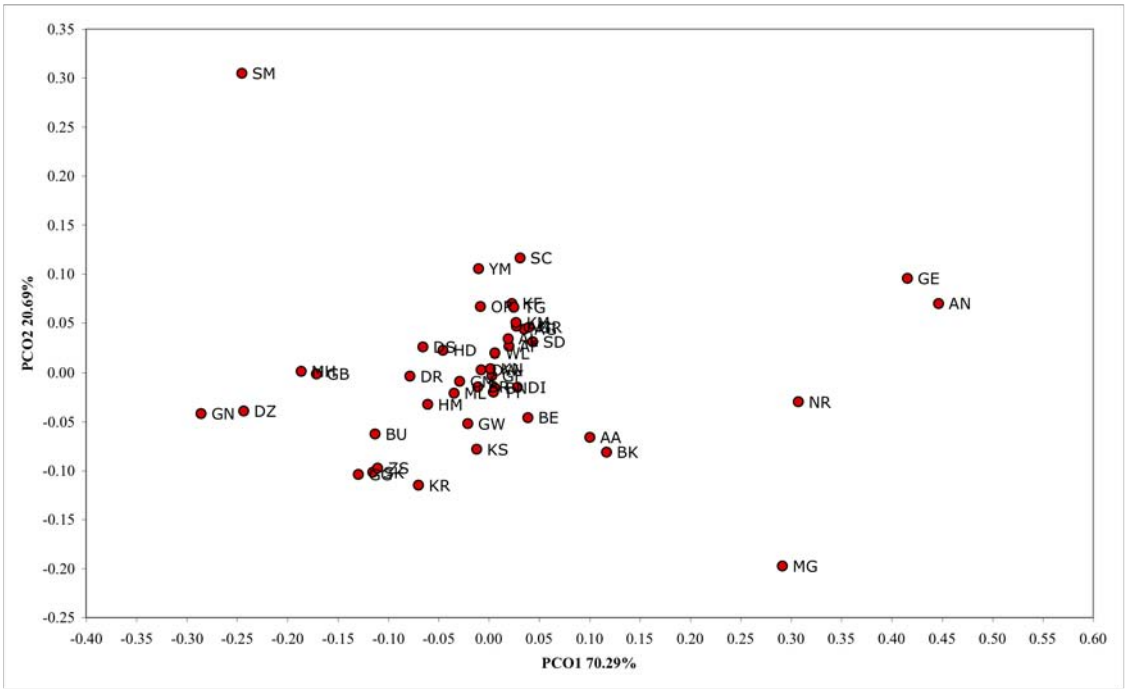
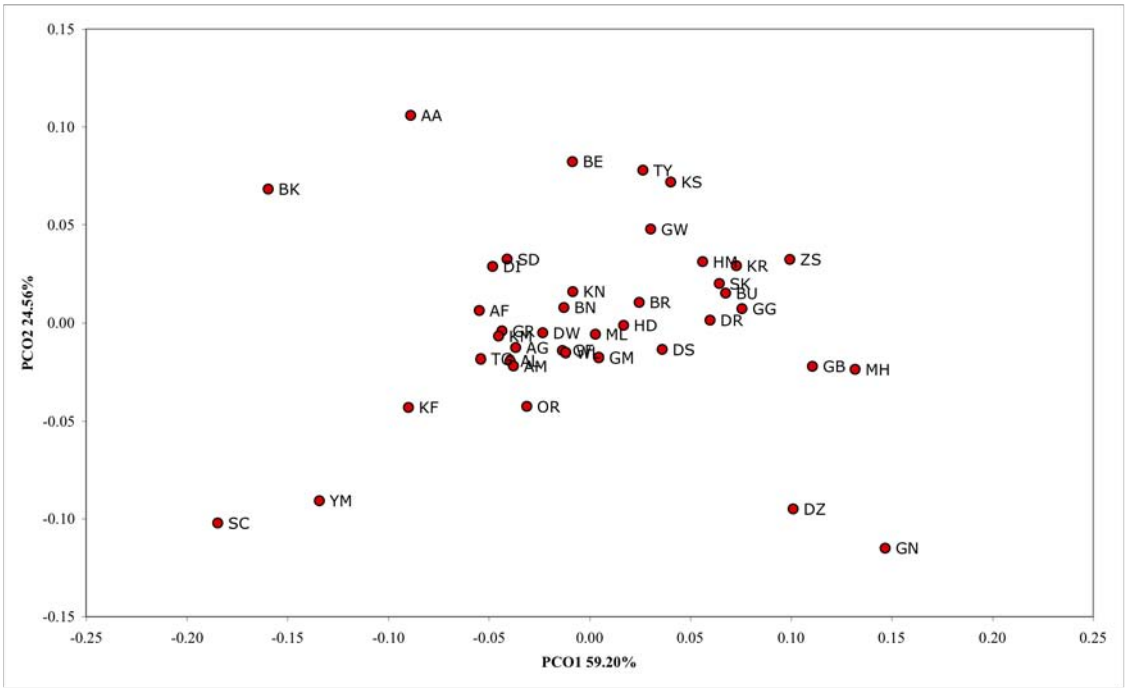


Figure 3.21 PCO of pairwise Rst distances between Ethiopian ethnic groups using NRY MS, excluding the Gedeo, Anuak, Nuer, Majenger and Somali



The PCO plot of pairwise Fst values using mtDNA HVS1 haplotypes (Figure 3.22) shows a diffuse cluster containing all the ethnic groups, apart from the Majenger, which appears as an outlier along both PCO1 and PCO2. The Majenger (MG) jointly with the Zayse (ZS) have the lowest *h* (estimated using HVS1 haplotype frequencies). Furthermore, the Majenger have by far the highest average number of samples per HVS1 haplotype (2.80 samples per haplotype, Figure 3.9).

Figure 3.22 PCO of pairwise Fst distances between Ethiopian ethnic groups using mtDNA HVS1 haplotypes

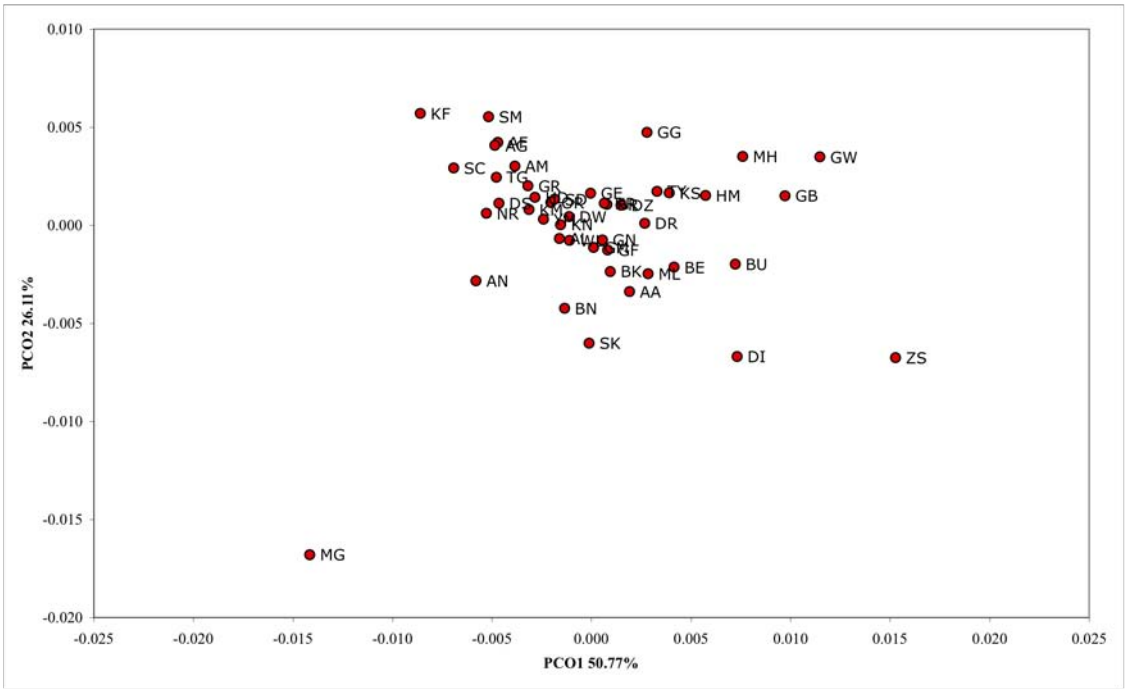
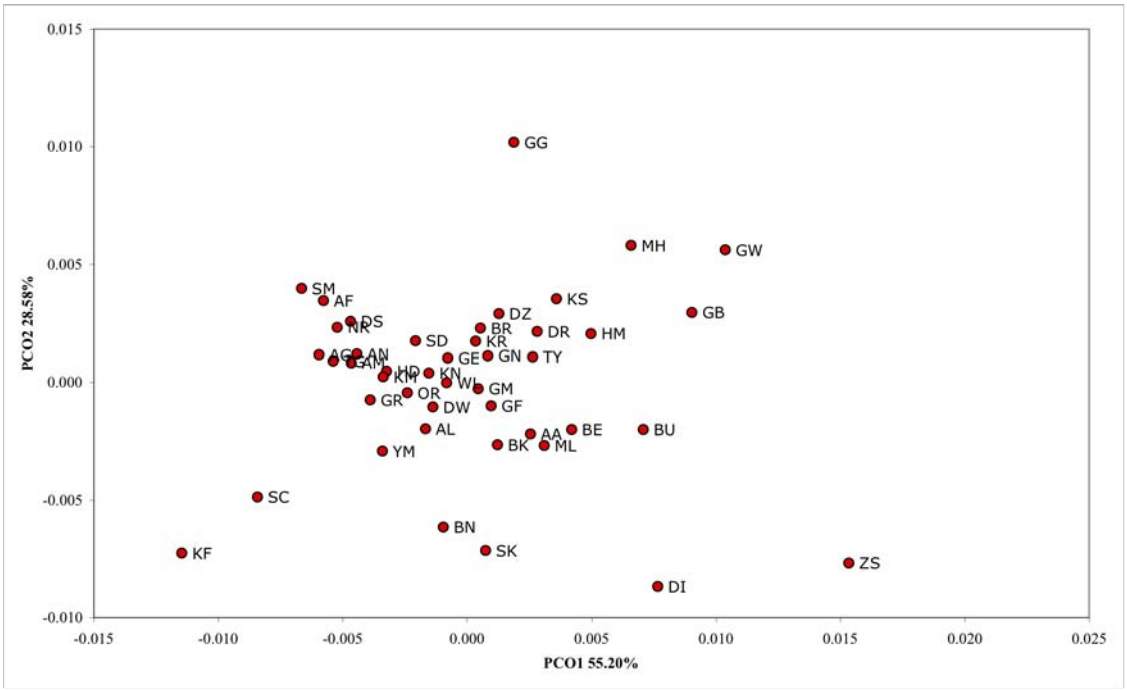


Figure 3.23 PCO of pairwise Fst distances between Ethiopian ethnic groups using mtDNA HVS1 haplotypes, excluding the Majenger



The PCO plot of K2P distances using mtDNA HVS1 haplotypes (Figure 3.24) shows a main cluster containing all the ethnic groups apart from the Majenger (MG), who appear as an outlier at the higher extremity of PCO1. The Somali (SM) and the Afar (AF) who appear together at the lower extremity of PCO1 are found in the low lying deserts of the south-east and north-east of Ethiopia respectively. The Agew (AG) and Tigray (TG) also appear together towards the lower extremity of PCO1. Both groups are found in the highlands of northern Ethiopia. The Nuer (NR) and Anuak (AN) appear as separate outliers along PCO2 (both groups are found in the Gambela province in western Ethiopia). After removal of the Majenger, Anuak, Nuer, Somali and Afar, some structuring within the main cluster becomes apparent (Figure 3.25). In particular, the Agew (AG), Tigray (TG) and Amhara (AM) appear together spread along the lower extremity of PCO1. There also seems to be a loose cluster of ethnic groups found in the northern area of the SNNP region (see section 2.4.3) and central Ethiopia, namely the Hadiya (HD), Kembata (KM), Konta (KN), Wolayta (WL), Dawuro (DW) and Oromo (OR), with the Gurage (GR) and Alaba (AL) appearing nearby. The outlier status of both the Dasanach (DS) and the Shekecho (SC) may be explained by their levels of nucleotide diversity, with the Dasanach possessing the highest π value, and the Shekecho possessing the lowest π value amongst all the ethnic groups surveyed.

Figure 3.24 PCO of pairwise K2P distances between Ethiopian ethnic groups using mtDNA HVS1 haplotypes

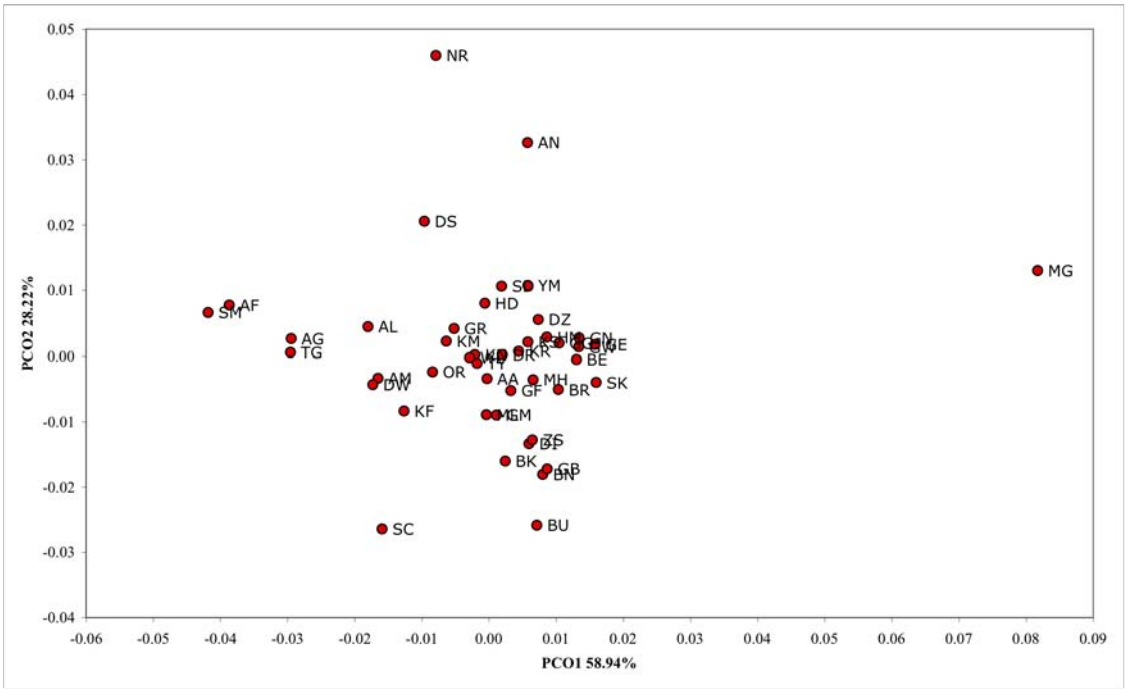
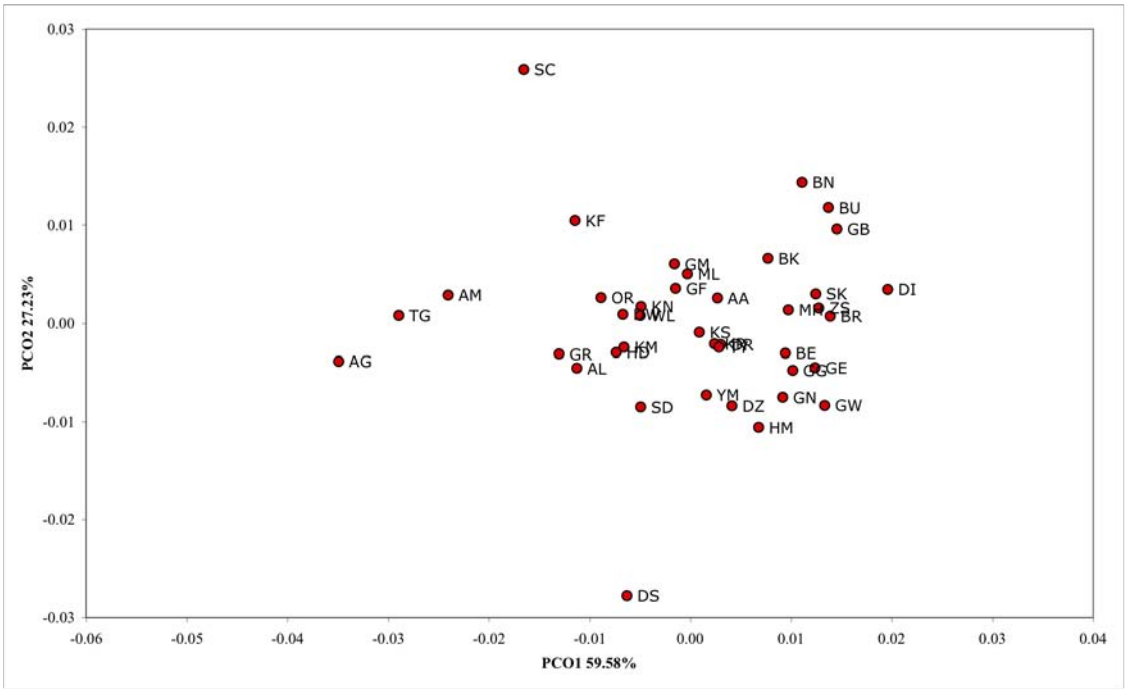


Figure 3.25 PCO of pairwise K2P distances between Ethiopian ethnic groups using mtDNA HVS1 haplotypes, excluding the Majenger, Anuak, Nuer, Afar and Somali



3.1.6 Genetic distances between Ethiopian ethnic groups and four non-Ethiopian groups

The first two principal coordinates determined for the pairwise genetic distances between Ethiopian ethnic groups, Igbo (IGB, green), Greek-Cypriots (CYP, blue), Fars (IRN, pink) and Halfawi (SUD, yellow) are shown in the PCO plots below. The complete table of all pairwise genetic distances can be found in Supplementary Table OtherDist, which was generated from the data in Supplementary Table OtherNRY and Supplementary Table OthermtDNA.

The PCO plot of pairwise Rst distances using MS haplotypes of the Ethiopian ethnic groups and the four other groups (Figure 3.26) is remarkably similar to the plot of Rst distances using only the Ethiopian groups (Figure 3.20), with the Igbo (IGB) appearing as the most distant outlier along PCO1, and the three other groups (the Fars (IRN), Greek-Cypriots (CYP) and Halfawi (SUD)) clustered together amongst the Ethiopian ethnic groups. After removal of the outlying Igbo (IGB), Majenger (MG), Anuak (AN), Gedeo (GE), Nuer (NR) and Somali (SM) (Figure 3.27), the remaining three global groups (Fars, Greek-Cypriots and Halfawi) appear at the higher extremity of PCO1, with the Kefa (KF), Yem (YM) and the Shekecho (SC) appearing nearby. This indicates the far greater similarity between many of the Ethiopian ethnic groups and those of West Asia (Fars), the Mediterranean (Greek-Cypriots) and North-East Africa (Halfawi) with respect to NRY variation, than with West Africa (Igbo). Additionally, this also shows the great deal of variation amongst the Ethiopian groups as seen in their wide distribution along both PCO1 and PCO2, which contrasts with the three non-Ethiopian groups (excluding the Igbo) which appear relatively clustered along PCO1 and PCO2 when compared to the distribution of Ethiopian ethnic groups.

Figure 3.26 PCO of pairwise Rst distances between Ethiopian ethnic groups and four other non-Ethiopian groups using NRY MS

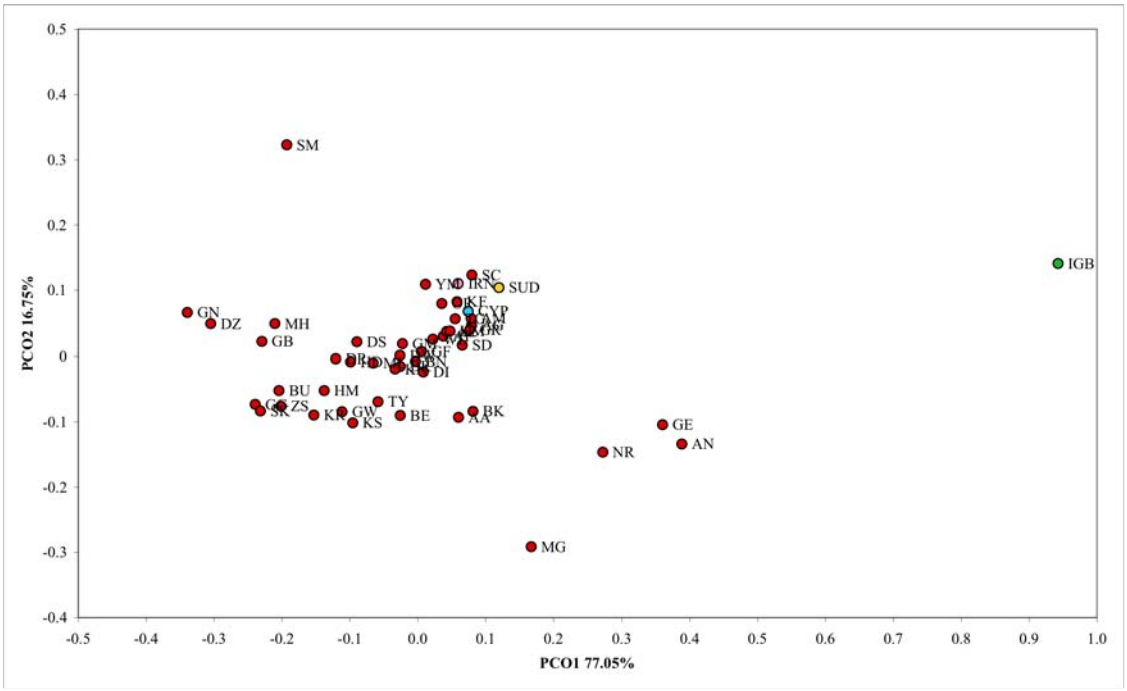
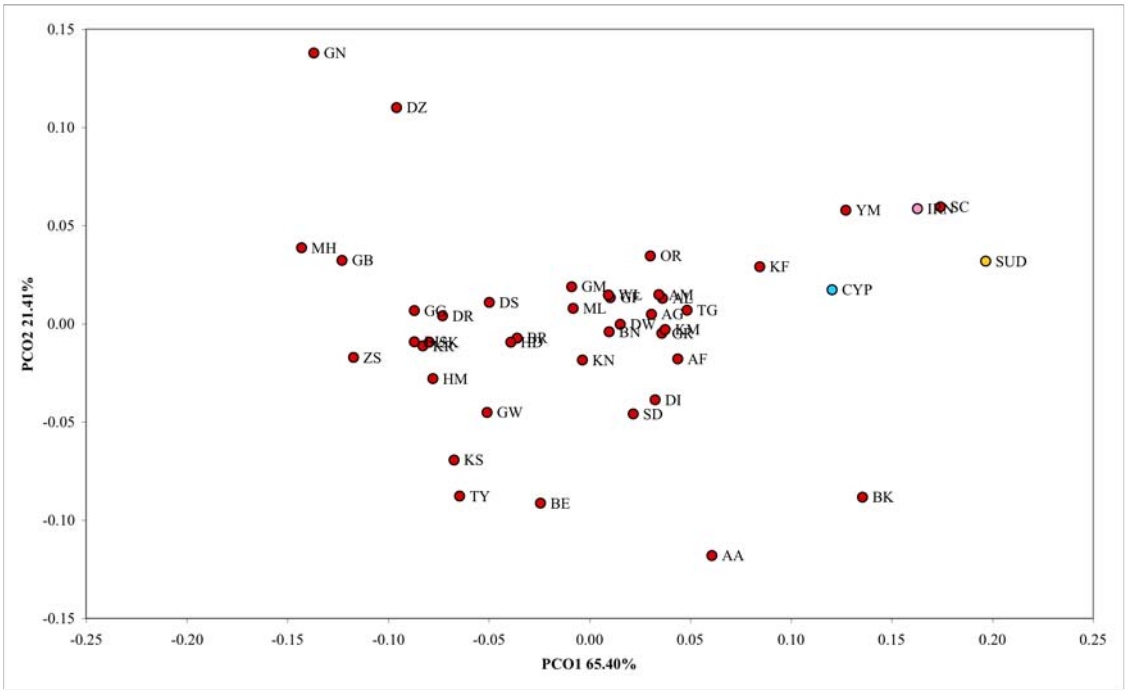


Figure 3.27 PCO of pairwise Rst distances between Ethiopian ethnic groups and three other non-Ethiopian groups using NRY MS, having excluded the Majenger, Anuak, Gedeo, Nuer, Somali and Igbo.



The PCO plot of K2P distances using mtDNA HVS1 haplotypes of the Ethiopian ethnic groups and the four other groups (Figure 3.28) shows a fairly diffuse cluster of the Ethiopian groups, with Fars (IRN) and Greek-Cypriots (CYP) appearing as clear outliers along PCO1. From Supplementary Table OtherDist it can be seen that the Fars and Greek-Cypriots show a non-significant pairwise K2P distance with each other, with all other distances highly significant. This could be partly explained by the high frequency of haplotypes identical to the Cambridge Reference Sequence in these two groups (20% in CYP, 26% in IRN), and very low frequency or absence in other groups. However, due to only 228bp of the mtDNA HVS1 sequence being used in this analysis, rather than the 381bp used in the rest of this thesis, it is likely that overall this has been responsible for greater similarity being shown amongst the groups due to the lower number of resolved haplotypes (824 compared with 1328 haplotypes resolved using 381bp of HVS1) than would otherwise have been the case. After removal of the Fars and Greek-Cypriots (Figure 3.29), the remaining groups show a diffuse spread along both PCO1 and PCO2. Interestingly, the Igbo (IGB), and the Halfawi (SUD) in particular appear nearby to some of the outlying Ethiopian groups, namely the Anuak (AN) and the Dasanach (DS), with the Nuer (NR) the furthest outlier along PCO1. These three Ethiopian ethnic groups were collected closest to the borders of present day Ethiopia (AN and NR near the western border, DS near the south-western border, see section 2.4.2 and 2.4.3), and are known to have large populations in neighbouring Sudan (AN and NR) and Kenya (DS) (www.ethnologue.com), and from these PCO plots they appear to be showing greater similarity with some African ethnic groups outside Ethiopia than with those within. Like the variation that is apparent amongst the plots of NRY Rst distances, these PCO plots of mtDNA K2P distances clearly demonstrate the substantial diversity that is present amongst the different ethnic groups in Ethiopia relative to the other groups included in this study.

Figure 3.28 PCO of pairwise K2P distances between Ethiopian ethnic groups and four other non-Ethiopian groups using mtDNA HVS1 haplotypes.

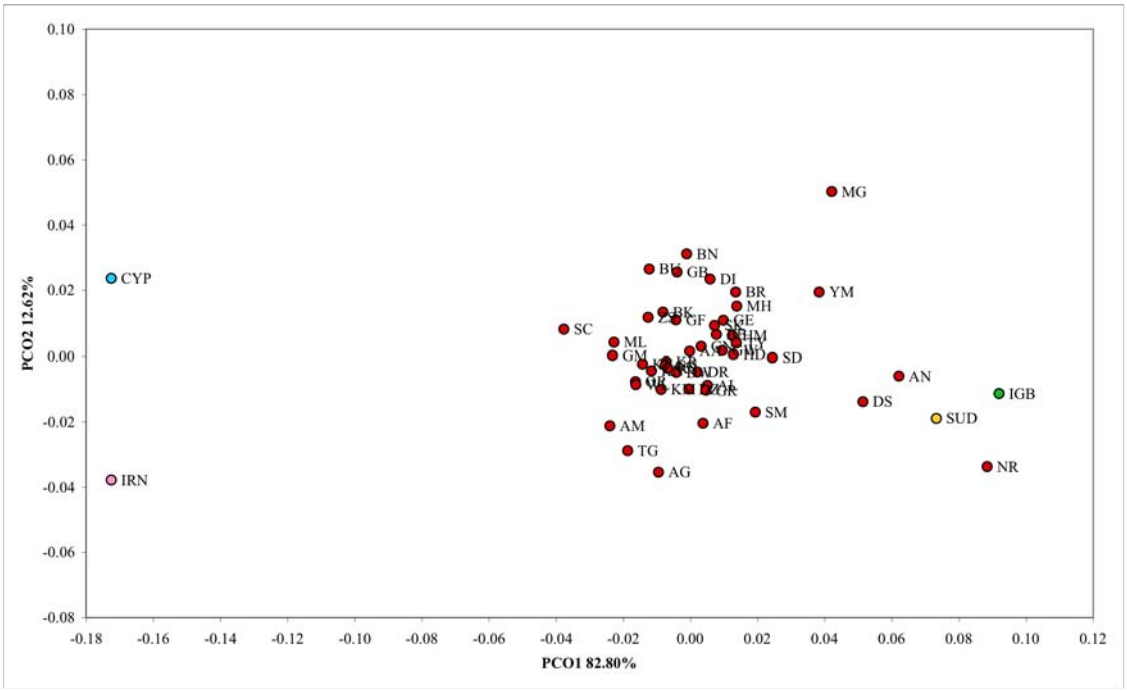
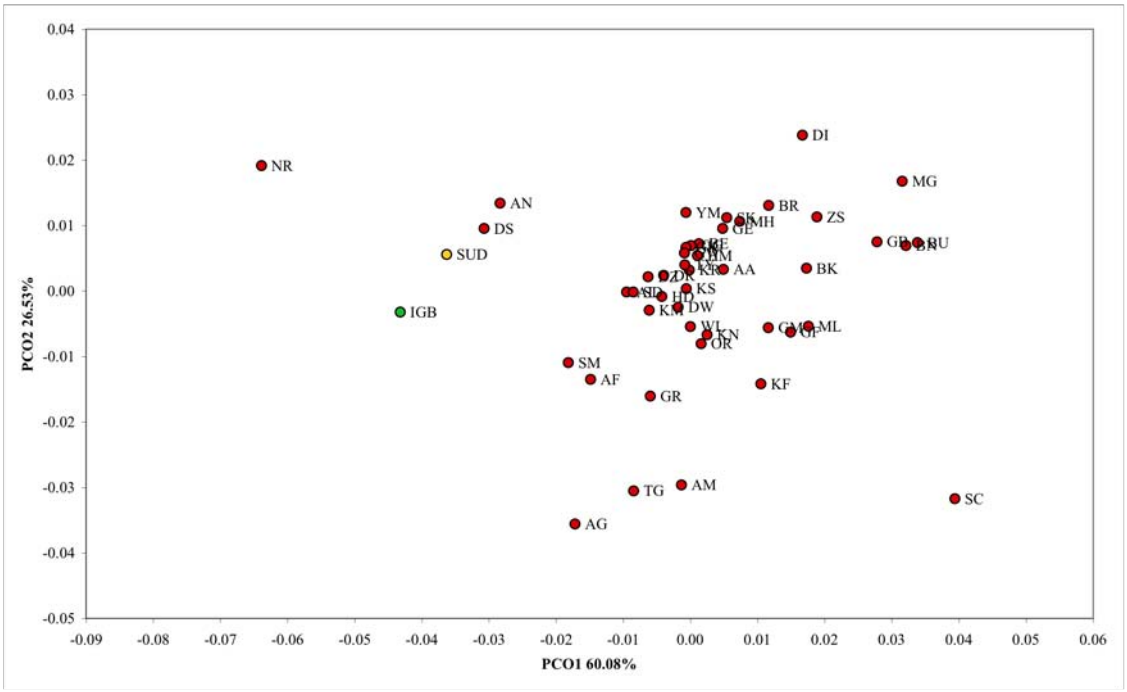


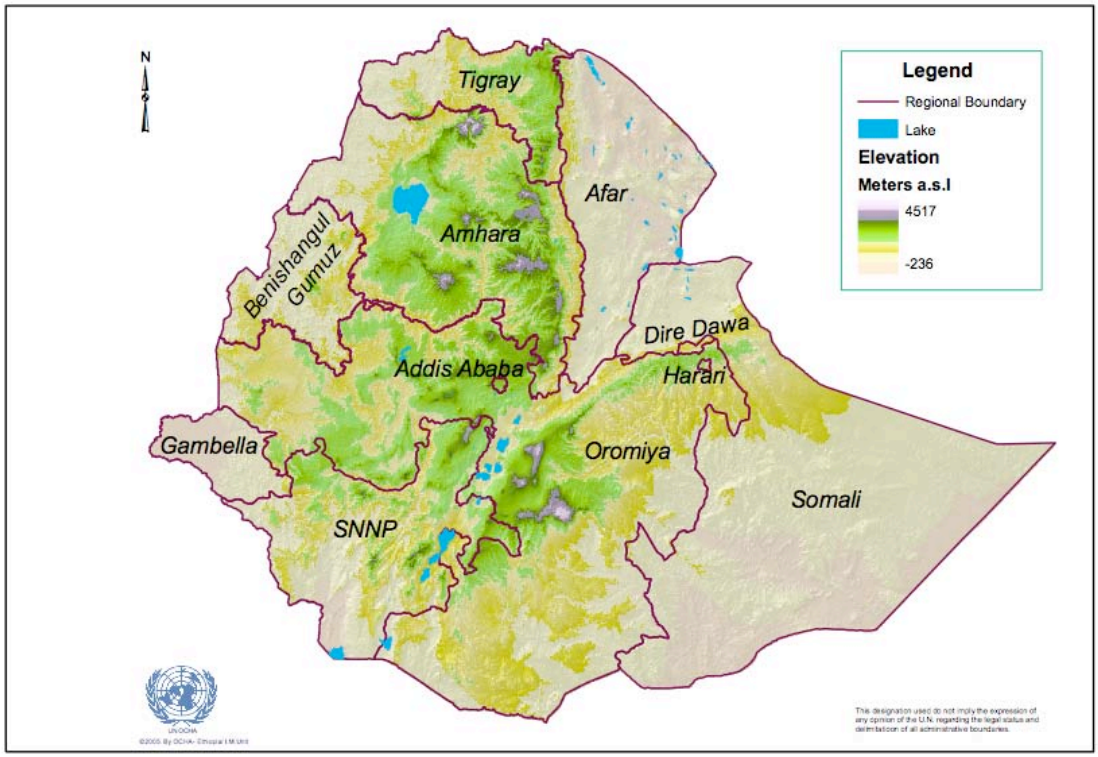
Figure 3.29 PCO of pairwise K2P distances between Ethiopian ethnic groups and two other non-Ethiopian groups using mtDNA HVS1 haplotypes, having excluded the Greek-Cypriots and Fars.



3.2 Given the ethnic basis of the modern administrative organisation of Ethiopia, can the people of the different provinces be differentiated from each other?

Note: It is appreciated that given the significant levels of variation among ethnic groups, comparisons between regions will be affected by the number of samples of each group represented in the analysis. However some guide to the interregional variation is provided by the results reported below.

Figure 3.30 Map of Ethiopian administrative regions (provinces)



Source: United Nations Office for the Coordination of Humanitarian Affairs: Ethiopia, http://www.ocha-eth.org/

78.1% of samples (4,497 of 5,756) were collected in the Southern Nations Nationalities and Peoples (SNNP) province (Table 3.3), which was the most ethnically diverse region (ethnic diversity 0.974). The Amhara province constitutes the second largest collection, but has a much lower ethnic diversity (0.541) as it is predominantly Amhara (AM) and Agew (AG) samples. Both the Afar province and the Chartered City of Dire Dawa (CC2) are ethnically uniform, with all samples belonging to the Afar (AF) and Somali (SM) ethnic groups respectively.

Table 3.3 Distribution of samples from ethnic groups collected in Ethiopian provinces

Ethnic group	Afar	Amhara	CC1	CC2	Gambela	Oromia	SNNP	Somali	Grand Total
AA							117		117
AF	112								112
AG		237	32						269
AL							110		110
AM		235	101		2	20	38		396
AN					108				108
BE							124		124
BK							113		113
BN							127		127
BR							119		119
BU							126		126
DI							132		132
DR							108		108
DS							105		105
DW							117		117
DZ							104		104
GB							113		113
GE							122		122
GF							111		111
GG							109		109
GM			1				208		209
GN							113		113
GR			24		7	2	119		152
GW							117		117
HD			1		1		125		127
HM							112		112
KF							120		120
KM			1		2		114		117
KN							107		107
KR							108		108
KS							120		120
MG							115		115
MH							130		130
ML							119		119
NR					118				118
OR			48		15	49	37		149
SC							125		125
SD							126		126
SK							113		113
SM				17			5	86	108
TG		21	11		3		30	1	66
TY							114		114
WL			2		1		107		110
YM						1	107		108
ZS							111		111
Total	112	493	221	17	257	72	4497	87	5756

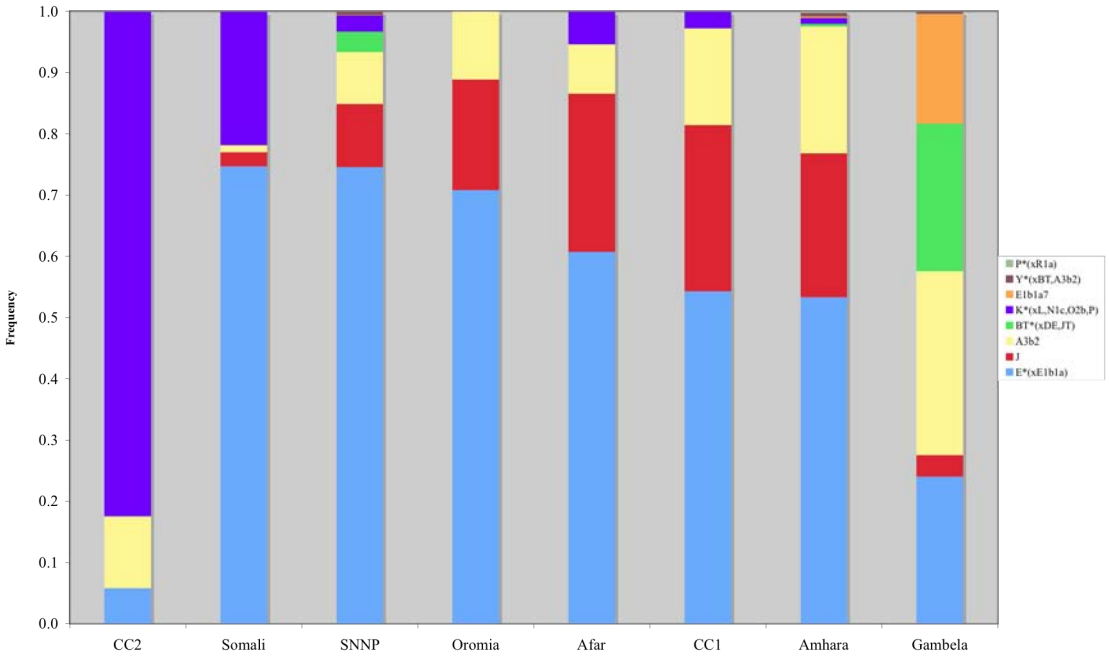
Table 3.4 Summary of the genetic diversity found within Ethiopian provinces

Province	NRY haplogroup *h*	s.d ±	NRY haplotype *h*	s.d ±	NRY MS MSV	mtDNA haplotype *h*	s.d ±	mtDNA haplotype π	s.d ±
Afar	0.560	0.047	0.940	0.022	0.851	0.987	0.011	0.0246	0.0126
Amhara	0.618	0.022	0.966	0.008	1.232	0.990	0.005	0.0233	0.0119
CC1	0.608	0.033	0.963	0.013	1.288	0.991	0.006	0.0240	0.0122
CC2	0.324	0.113	0.596	0.119	0.550	0.985	0.029	0.0236	0.0128
Gambela	0.764	0.027	0.957	0.013	0.931	0.993	0.005	0.0262	0.0133
Oromia	0.460	0.059	0.962	0.022	1.058	0.991	0.011	0.0250	0.0128
SNNP	0.424	0.007	0.976	0.002	0.832	0.994	0.001	0.0244	0.0124
Somali	0.398	0.052	0.895	0.033	0.817	0.980	0.015	0.0225	0.0116

3.2.1 NRY diversity

Haplogroup level gene diversity (*h,* Table 3.4) in provinces ranged from 0.764 in Gambela, to 0.324 in Dire Dawa (CC2). The mean *h* value across all provinces was 0.519, and the median value 0.510. The Gambela province is notable in having both the highest frequency of haplogroups A3b2 and BT*(xDE,JT) (30.0% and 24.1% respectively), and the only region in Ethiopia with a substantial frequency of haplogroup E1b1a7 (17.9% frequency). Dire Dawa is notable in that 82.4% of samples belong to haplogroup K*(xL,N1c,O2b,P), a far higher frequency than that found in any other province. Gene diversity using frequencies of UEP-MS haplotypes ranged from 0.976 in SNNP to 0.596 in Dire Dawa, with mean and median *h* values of 0.907 and 0.959 respectively. MSV ranged from 1.288 in Addis Ababa (CC1) to 0.550 in Dire Dawa, with mean and median MSV of 0.945 and 0.891 respectively. Interestingly, the Somali province, which borders the chartered city of Dire Dawa in eastern Ethiopia, displayed a similar pattern of diversity to Dire Dawa in that it exhibited the second lowest NRY diversity values (after Dire Dawa) as estimated by all three metrics.

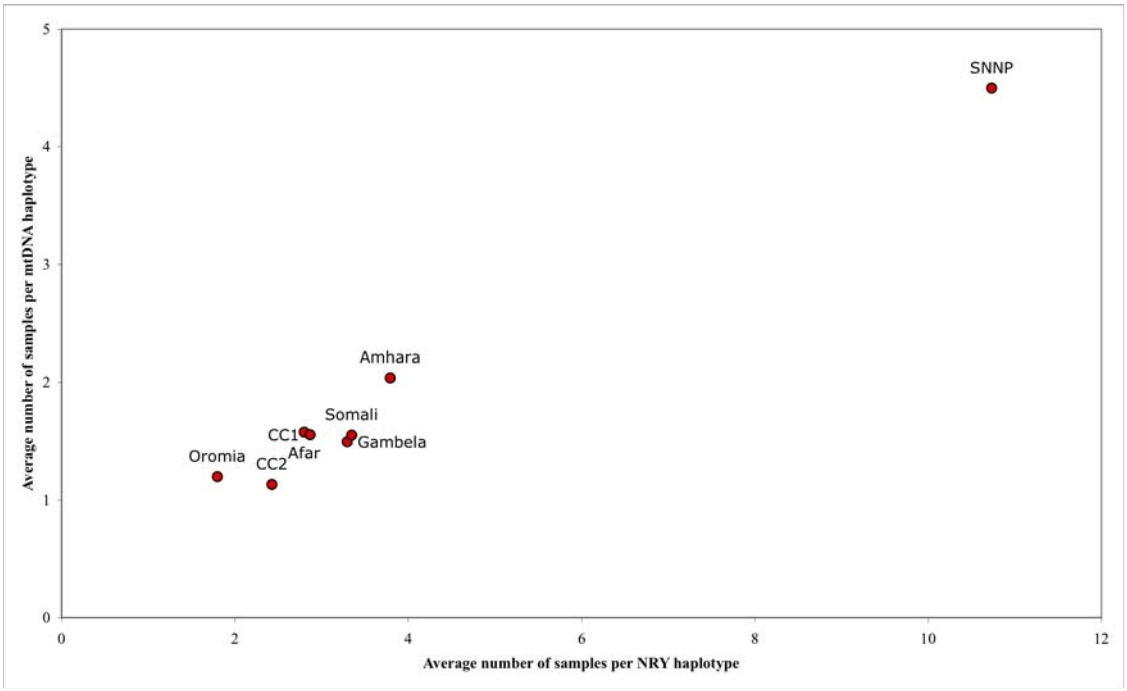
Figure 3.31 Frequencies of NRY haplogroups found in Ethiopian provinces (ordered by increasing haplogroup gene diversity)



3.2.2 mtDNA diversity

Gene diversity (Table 3.4) estimated using mtDNA HVS1 haplotype frequencies in provinces ranged from 0.994 in the SNNP, to 0.980 in the Somali province, with mean and median *h* values of 0.989 and 0.991 respectively. Both the mean and median nucleotide diversity (π) was 0.0242, with values ranging from 0.0262 in Gambela to 0.0225 in the Somali. The Gambela has the second highest haplotype level diversity, and by far the highest nucleotide diversity, indicating the presence of high numbers of low frequency divergent haplotypes in the Gambela, whereas the haplotypes found in the Somali province tend to be of a higher frequency and are more similar to each other.

Figure 3.32 Plot of the average number of samples per haplotype per province for NRY and mtDNA



3.2.3 Correlation between NRY and mtDNA diversity

A significant linear correlation was observed between the numbers of samples per NRY haplotype and the numbers of samples per mtDNA HVS1 haplotype in provinces (r^2=0.9828, p<0.0001, Figure 3.32). For a given province, there was approximately twice the number of samples per NRY haplotype compared to the number of samples per mtDNA haplotype, with mean values across all provinces of 3.883 and 1.881 per dataset respectively. There were moderately significant correlations (0.05>p>0.01, Table 3.5) between NRY UEP haplogroup gene diversity values and NRY MSV values, as well as NRY UEP-MS haplotype gene diversity values and mtDNA HVS1 haplotype gene diversity values. All other correlations between diversity values were not significant, possibly due to the small number of data points.

Table 3.5 Correlations between ranked diversity values in provinces (p-values in upper diagonal, correlation coefficients in lower diagonal)

	NRY UEP *h*	NRY UEP-MS *h*	NRY MS MSV	mtDNA HVS1 *h*	mtDNA HVS1 π
NRY UEP *h*	*	0.2675	**0.0279**	0.2431	0.2992
NRY UEP-MS *h*	0.4524	*	0.1150	**0.0368**	0.7520
NRY MS MSV	0.7857	0.6190	*	0.2992	0.5821
mtDNA HVS1 *h*	0.4791	0.7545	0.4311	*	0.0831
mtDNA HVS1 π	0.4286	0.1429	0.2381	0.6587	*

3.2.4 Exact Tests of Population Differentiation between provinces

ETPD using NRY UEP haplogroup frequencies (Table 3.6) showed Dire Dawa (CC2), Gambela and the Somali Province to be distinct from all other provinces ($p<0.01$ for all comparisons). These three provinces have high frequencies of UEP haplogroups that are rare in the other provinces, namely E1b1a7 (in Gambela) and K*(xL,N1c,O2b,P) (in Dire Dawa and the Somali province). The Oromia province was the least differentiated from other provinces, with non-significant differences shown between Oromia and the Afar, Amhara, Addis Ababa (CC1) and SNNP provinces. Additionally, Addis Ababa was not significantly differentiated from both the Afar and Amhara provinces, although these two provinces were distinct from each other.

Table 3.6 Matrix of ETPD p values based on NRY UEP haplogroup frequencies in provinces ($p>0.01$ in bold)

	Afar	Amhara	CC1	CC2	Gambela	Oromia	SNNP	Somali
Afar	*							
Amhara	0.003	*						
CC1	**0.137**	**0.308**	*					
CC2	<0.001	<0.001	<0.001	*				
Gambela	<0.001	<0.001	<0.001	<0.001	*			
Oromia	**0.099**	**0.250**	**0.072**	<0.001	<0.001	*		
SNNP	<0.001	<0.001	<0.001	<0.001	<0.001	**0.134**	*	
Somali	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	*

ETPD performed using NRY UEP-MS haplotype frequencies (Table 3.7) showed the Amhara province to be non-differentiated from both Addis Ababa (CC1) and Oromia. Additionally, Dire Dawa (CC2) was moderately differentiated ($0.05>p>0.01$) from the Somali province. The Afar, Gambela and SNNP provinces were shown to be significantly differentiated from all other provinces ($p<0.01$ for all comparisons).

Table 3.7 Matrix of ETPD p values based on NRY UEP-MS haplotype frequencies in provinces (p>0.01 in bold)

	Afar	Amhara	CC1	CC2	Gambela	Oromia	SNNP	Somali
Afar	*							
Amhara	<0.001	*						
CC1	<0.001	**0.112**	*					
CC2	<0.001	<0.001	<0.001	*				
Gambela	<0.001	<0.001	<0.001	<0.001	*			
Oromia	<0.001	**0.145**	0.006	<0.001	<0.001	*		
SNNP	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	*	
Somali	<0.001	<0.001	<0.001	**0.017**	<0.001	<0.001	<0.001	*

Using the frequencies of mtDNA HVS1 haplotypes, ETPD showed the Afar and Gambela provinces to be distinct from all others ($p<0.01$ for all comparisons, Table 3.8. Dire Dawa (CC2) was shown to be moderately non-differentiated from both the Amhara and Oromia provinces ($0.05>p>0.01$), and not differentiated from the Somali province ($p=0.102$). Non-significant differences were also shown between Addis Ababa (CC1) and both the Amhara and Oromia provinces, although these provinces were shown to be distinct from each other.

Table 3.8 Matrix of ETPD p values based on mtDNA HVS1 haplotype frequencies in provinces (p>0.01 in bold)

	Afar	Amhara	CC1	CC2	Gambela	Oromia	SNNP	Somali
Afar	*							
Amhara	<0.001	*						
CC1	<0.001	**0.190**	*					
CC2	0.001	**0.020**	0.004	*				
Gambela	<0.001	<0.001	<0.001	<0.001	*			
Oromia	<0.001	0.002	**0.116**	**0.048**	<0.001	*		
SNNP	<0.001	<0.001	<0.001	<0.001	<0.001	**0.052**	*	
Somali	<0.001	<0.001	<0.001	**0.102**	<0.001	<0.001	<0.001	*

3.2.5 Genetic distances between provinces

The first two principal coordinates determined for the pairwise genetic distances between Ethiopian provinces are shown in the PCO plots below. Due to the small sample size of Dire Dawa (CC2), and its low diversity using all metrics (Table 3.4), PCO was often performed both with and without the inclusion of this province. The complete table of all pairwise genetic distances can be found in Supplementary Table ProvDist.

The PCO plot of pairwise Fst using UEP haplogroup frequencies (Figure 3.33) shows the provinces spread along PCO1, with Dire Dawa as an outlier at the lower extremity. Dire Dawa has by far the highest frequency of haplogroup K*(xL,N1c,O2b,P) (82.4%, Figure 3.31), and consequently by far the lowest frequency of haplogroup E*(xE1b1a), the modal haplogroup in all provinces except Gambela. After removal of Dire Dawa (Figure 3.34), the PCO plot shows Gambela as an outlier at the higher extremity of PCO1, and SNNP at the lower extremity. The Gambela province has by far the highest UEP haplogroup level *h* value (Table 3.4), as well as the lowest frequency of haplogroup E*(xE1b1a), whereas the SNNP and Somali provinces have the two lowest haplogroup level *h* after Dire Dawa, and the two highest frequencies of haplogroup E*(xE1b1a). The Somali province has the highest frequency of haplogroup K*(xL,N1c,O2b,P) (21.8%) after Dire Dawa, a haplogroup which is at less than 6% frequency in all other provinces, and the lowest frequency of haplogroup A3b2 (1.1%), which is at greater than 8% frequency in all other provinces.

Figure 3.33 PCO of pairwise Fst distances between Ethiopian provinces using NRY UEP haplogroup frequencies

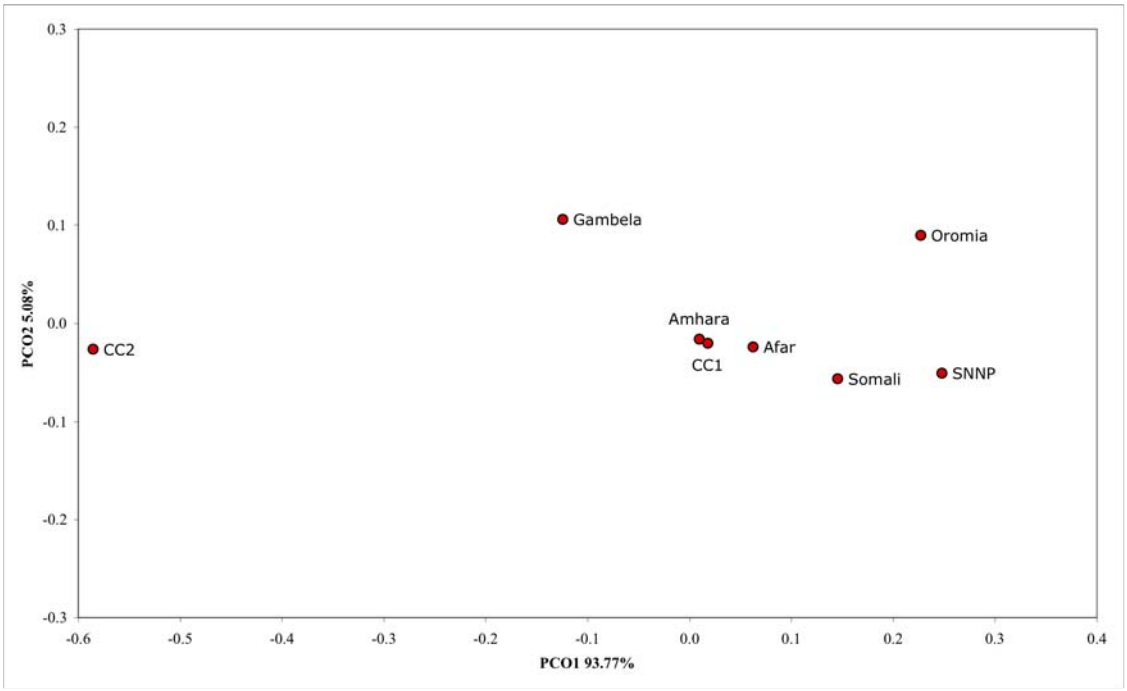
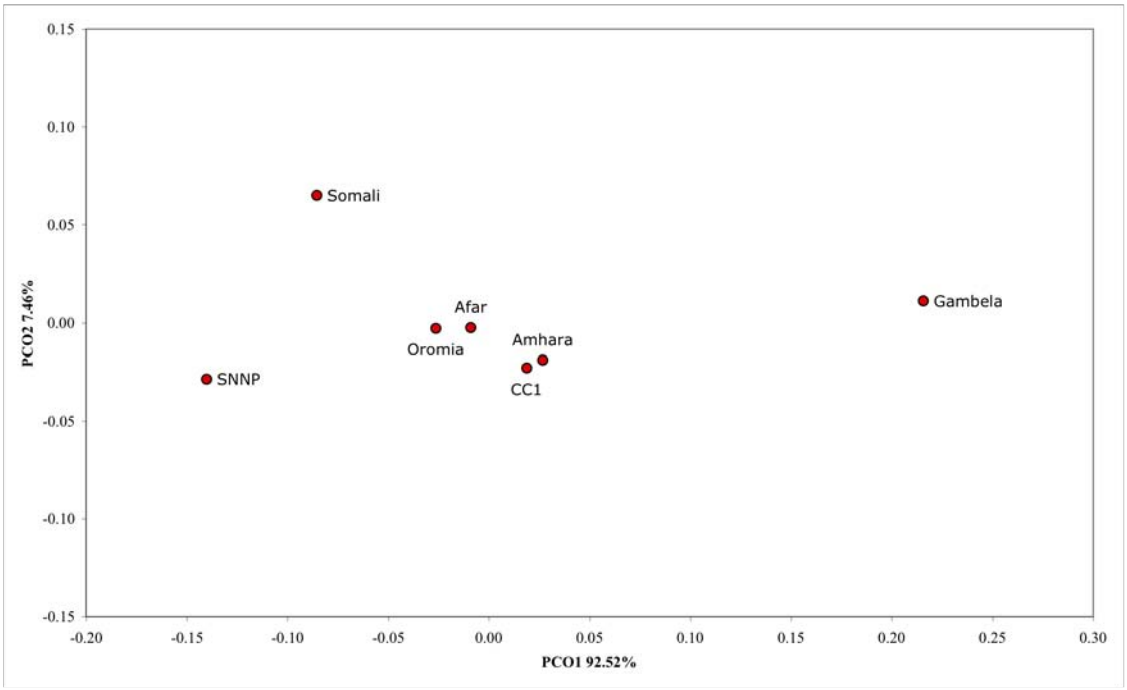


Figure 3.34 PCO of pairwise Fst distances between Ethiopian provinces using NRY UEP haplogroup frequencies, excluding Dire Dawa



The PCO plot of pairwise Fst using NRY UEP-MS haplotypes (Figure 3.35), shows Dire Dawa (CC2) as an outlier at the higher extremity of PCO1, with all other provinces appearing as a cluster at the lower extremity. Dire Dawa has by far the lowest *h* value based on UEP-MS haplotype frequencies (Table 3.4, *h*=0.596, mean *h*=0.907), compared to all other provinces. After removal of Dire Dawa (Figure 3.36), the Somali province appears as an outlier at the higher extremity of PCO1, and the Afar and Gambela provinces towards the lower extremity. Gambela and the Afar province also appear as outliers along PCO2 at the higher and lower extremity respectively. The Somali, Afar and Gambela provinces have respectively the lowest *h* values based on UEP-MS haplotype frequencies, after Dire Dawa.

Figure 3.35 PCO of pairwise Fst distances between Ethiopian provinces using NRY UEP-MS haplotype frequencies

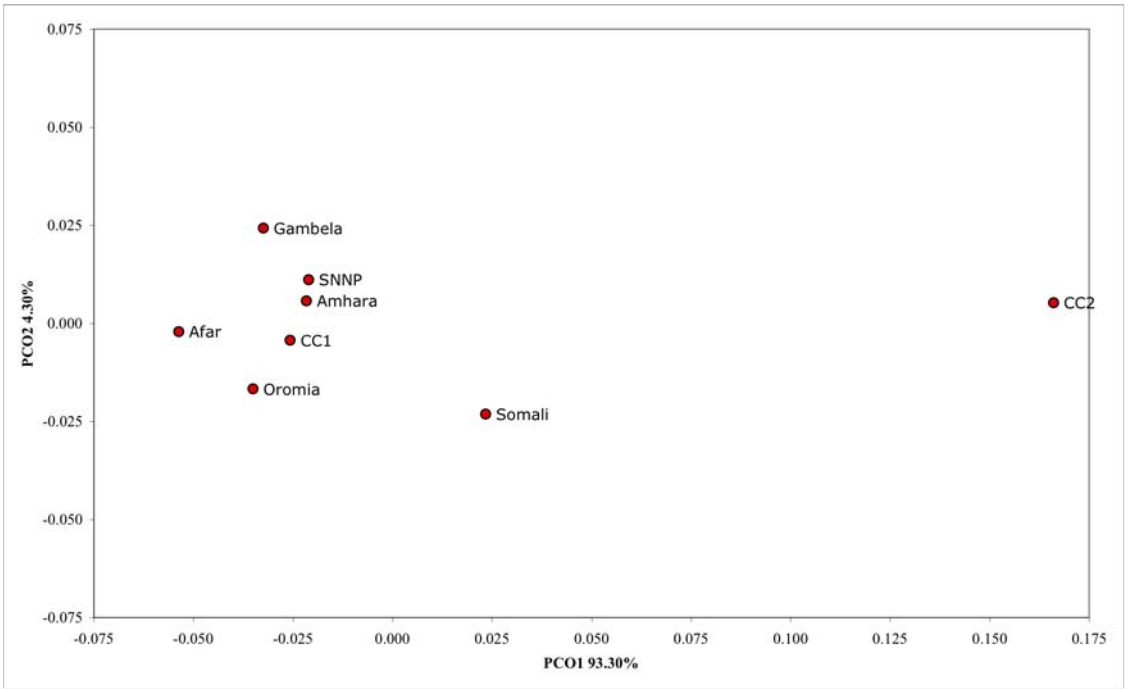
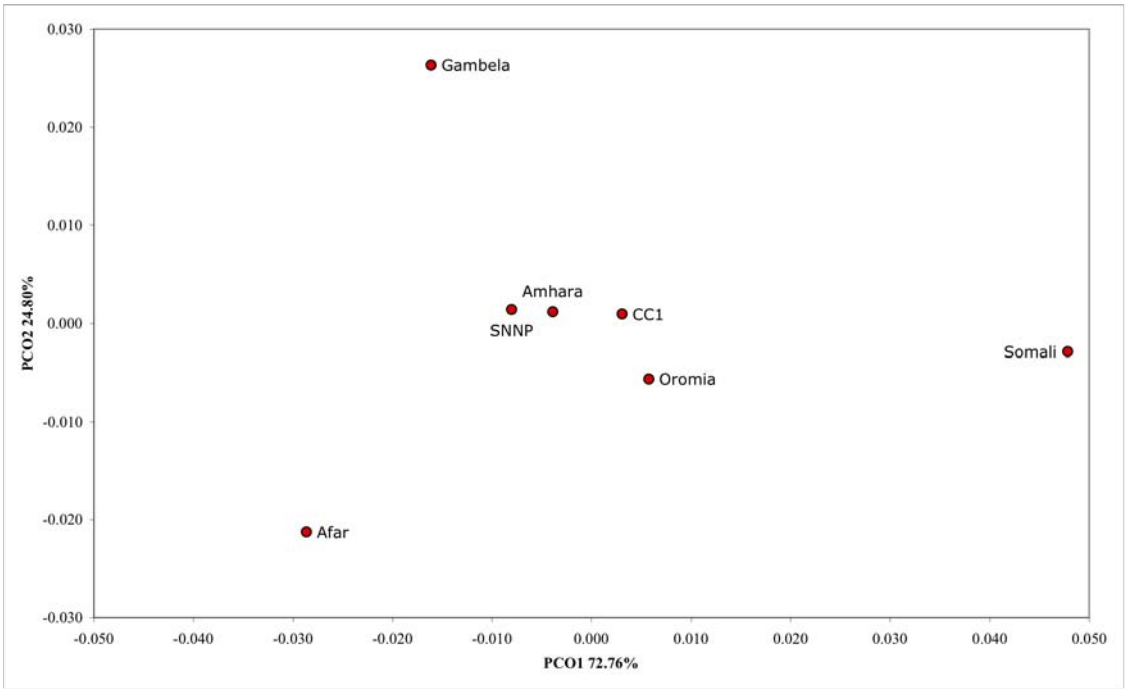


Figure 3.36 PCO of pairwise Fst distances between Ethiopian provinces using NRY UEP-MS haplotype frequencies, excluding Dire Dawa



The PCO plot of pairwise Rst distances using NRY MS (Figure 3.37) shows a central cluster containing 5 of the 8 provinces, with the Somali province appearing as an outlier at the higher extremity of PCO1 and Gambela appearing as an outlier at the lower extremity, whereas Dire Dawa (CC2) appears as an outlier at both the lower extremity of PCO1 and PCO2. Dire Dawa has by far the lowest MSV value (Table 3.4, MSV=0.550, mean MSV=0.945) and the highest frequency of a single haplogroup (82.4% K*(xL,N1c,O2b,P)), compared to all other provinces. After removal of Dire Dawa (Figure 3.38), the PCO plot of the remaining ethnic groups appear along PCO1, with Gambela as an outlier at the higher extremity, and the Somali province at the lower extremity. All Rst distance p values were significant ($p<0.01$, Supplementary Table ProvDist), except between Addis Ababa (CC1) and the Amhara and Oromia provinces ($p=0.052$ and $p=0.047$ respectively). Gambela has the highest frequencies of haplogroups BT*(xDE,JT) (24.1%) and E1b1a7 (17.9%), which are relatively rare in the rest of Ethiopia (absent or at less than 5% frequency in all other provinces), and due to the low level of microsatellite haplotypes shared among haplogroups (4.2% of haplotypes), this would be expected to cause substantial Rst distances between Gambela and the other provinces. Likewise, the Somali province has the highest frequency of haplogroup K*(xL,N1c,O2b,P) after Dire Dawa (21.8%), but has also the lowest frequency of haplogroup A3b2 (1.1% frequency, at greater than 8% frequency in all

other provinces). Additionally, the Somali province has the lowest NRY diversity after Dire Dawa (Table 3.4), as determined by both haplotype and haplogroup level gene diversity, and mean microsatellite repeat length variance, indicating that the haplotypes present in the Somali province are generally more clustered due to their similarity than those found in other provinces.

Figure 3.37 PCO of pairwise Rst distances between Ethiopian provinces using NRY MS

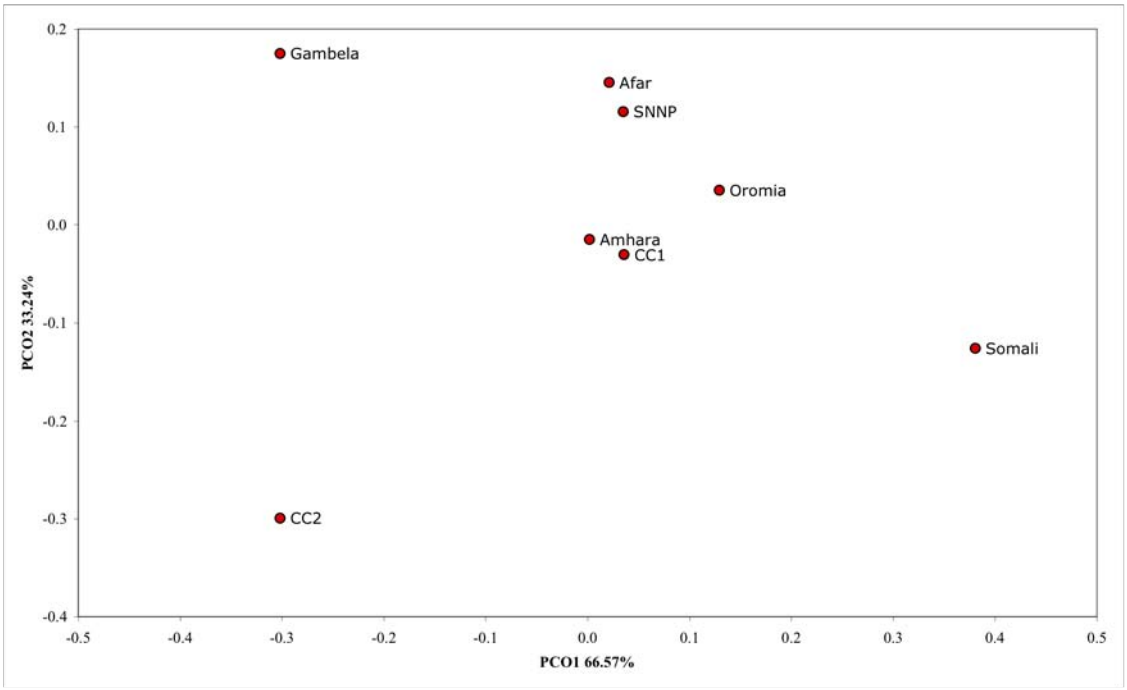
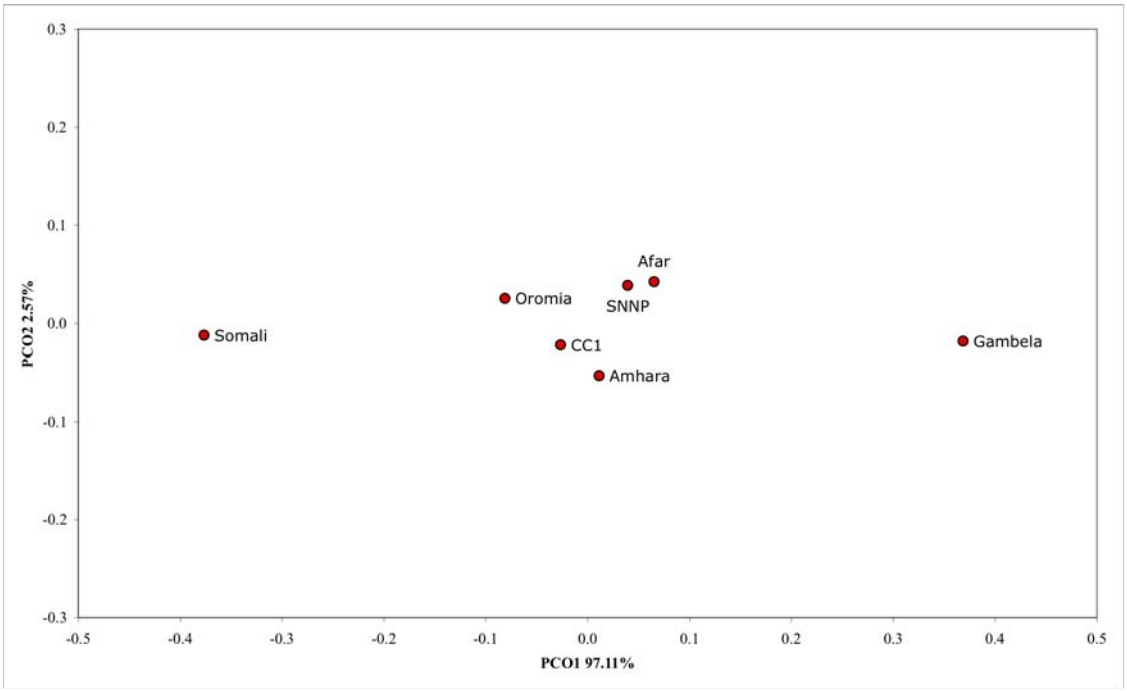


Figure 3.38 PCO of pairwise Rst distances between Ethiopian provinces using NRY MS, excluding Dire Dawa



The PCO plot of pairwise Fst values using mtDNA HVS1 haplotypes (Figure 3.39), shows the Somali province and Dire Dawa (CC2) as outliers at higher extremity of PCO1 and all other provinces appearing at the lower extremity. The Somali province and Dire Dawa have the two lowest mtDNA HVS1 haplotype level *h* values (Table 3.4, *h*=0.980 and *h*=0.985 respectively, mean *h*=0.989). After removal of Dire Dawa and the Somali province (Figure 3.40), Gambela and the Afar province appear as outliers at the lower and higher extremity of PCO1 respectively, and also both appear as outliers at the higher extremity of PCO2. The distribution of provinces along PCO1 may be due to their geographical location, as the order of provinces conforms to a rough north-east (Afar) to south-west (Gambela) transect.

Figure 3.39 PCO of pairwise Fst distances between Ethiopian provinces using mtDNA HVS1 haplotype frequencies

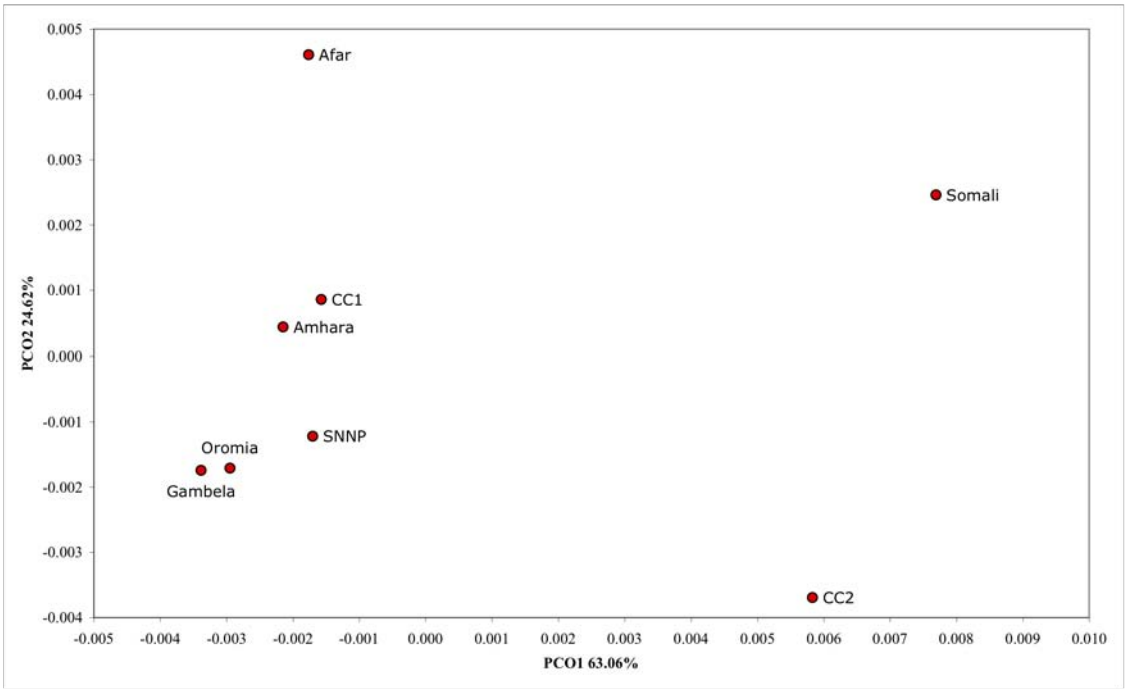
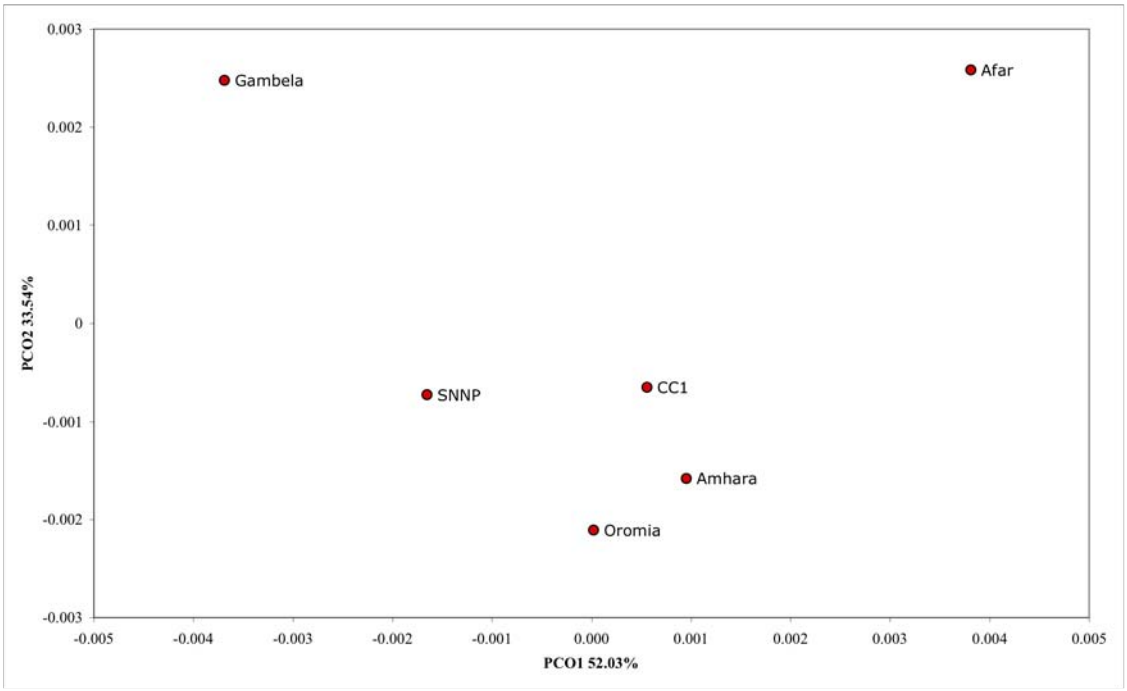


Figure 3.40 PCO of pairwise Fst distances between Ethiopian provinces using mtDNA HVS1 haplotype frequencies, excluding Dire Dawa and the Somali province



The PCO plot of K2P distances using mtDNA HVS1 haplotypes (Figure 3.41) show Dire Dawa (CC1) appearing as an outlier at the lower extremity of PCO1, and the SNNP province and Gambela appear as outliers at the higher extremity. Along PCO2, Gambela is an outlier at the lower extremity, whereas all the other provinces are clustered with the Amhara province at the higher extremity. After removal of Dire Dawa (Figure 3.42), Gambela now appears as an outlier at the higher extremity of PCO1, with the Amhara province appearing at the lower extremity. Interestingly, the Afar and Somali provinces appear clustered together on both PCO1 and PCO2. When the geographic positions of these two provinces are compared with their positions on the PCO plot of pairwise Fst distances (Figure 3.39), it becomes apparent that although the distribution of haplotypes present in these two provinces are substantially different from each other, there is still a greater degree of similarity between their haplotypes, than either province displays with haplotypes present in other provinces. Similar to the distribution of the provinces on the plot of pairwise Fst distances, the order of the provinces along PCO1 of the plot of K2P distances (Figure 3.42) may be reflecting their geographical position, with the northern and eastern provinces appearing towards the lower extremity, and the southern and western provinces appearing toward the higher extremity.

Figure 3.41 PCO of pairwise K2P distances between Ethiopian provinces using mtDNA haplotypes

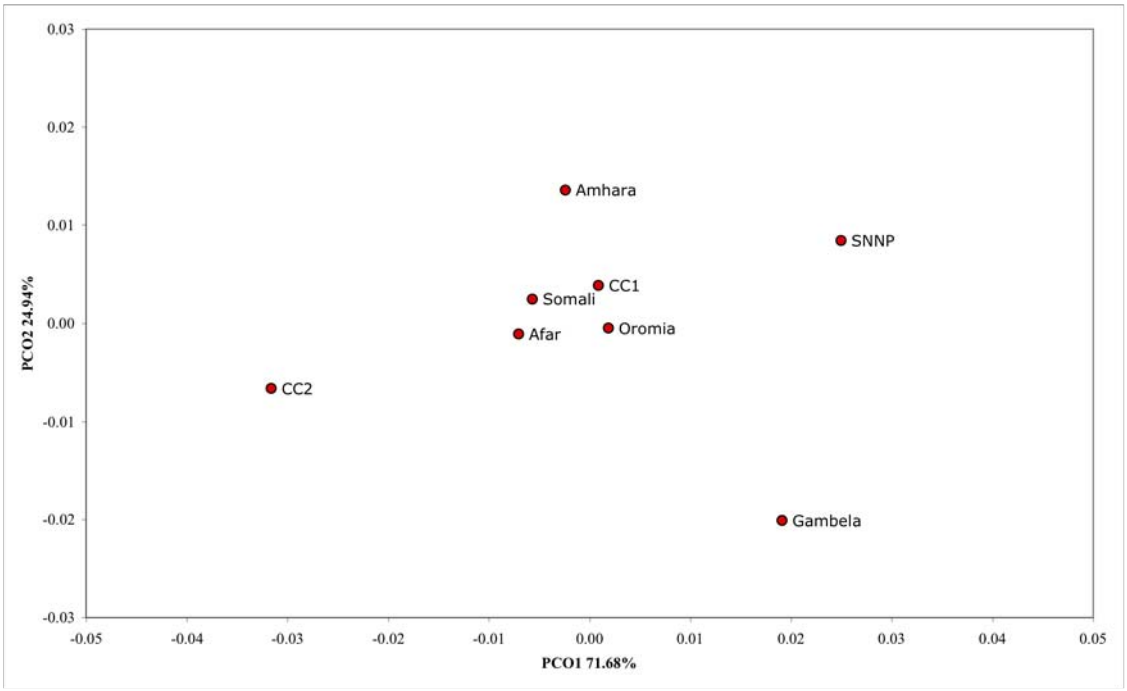
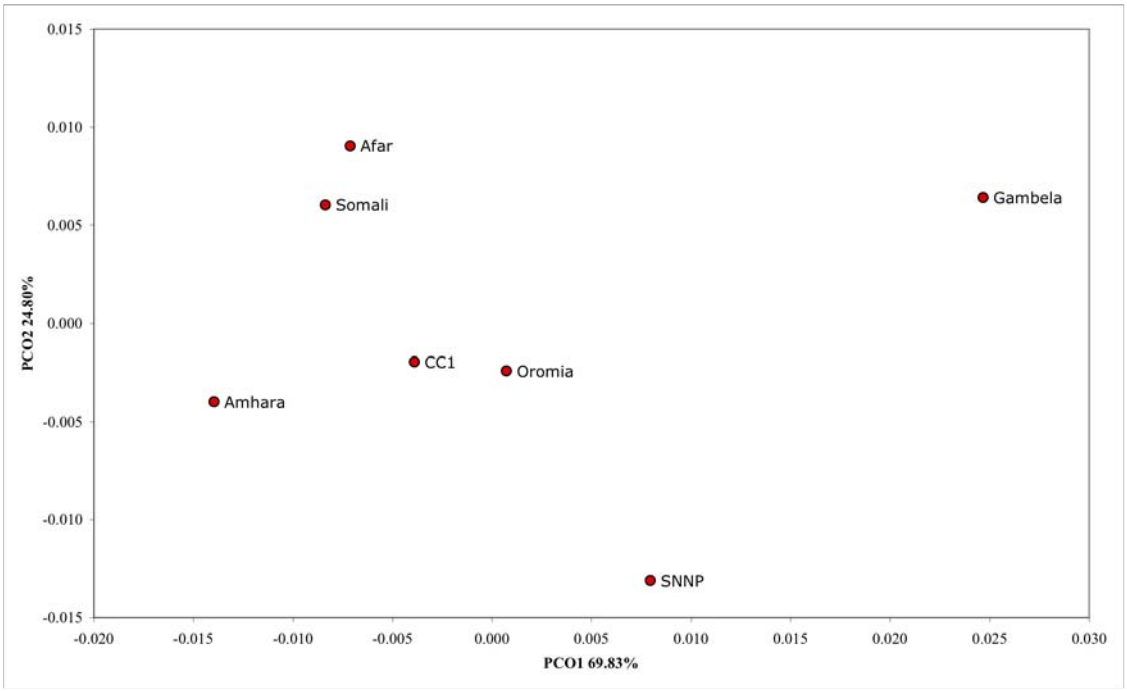


Figure 3.42 PCO of pairwise K2P distances between Ethiopian provinces using mtDNA haplotypes, excluding Dire Dawa



3.2.6 AMOVA to assess the general geographical apportionment of variance

The general geographic structure of the apportionment of variance was assessed hierarchically using Analysis of Molecular Variance (AMOVA, (Excoffier et al. 1992)), by grouping ethnic groups according to their traditional geographic regions of residence. These regions were Northern (N) consisting of Amhara, Agew and Tigray, Eastern (E) consisting of Afar and Somali, Western (W) consisting of Anuak and Nuer, and Southern (S) consisting of the remaining ethnic groups, traditionally resident in the Oromia and SNNP provinces.

For all metrics, the variance attributable to differences among ethnic groups (Global Fst), as well as differences attributable to differences among ethnic groups within geographic regions (Fsc), was highly significant ($p<0.001$, Table 3.9). Additionally, the proportion of variance attributable to differences among the geographic regions (Fct) was significant ($p<0.01$) for all metrics with the exception of NRY UEP-MS Fst. As Veeramah et al. (2010) established, Fst calculated using highly characterised sex-specific haplotypes can increase rapidly, over a few generations, in the absence of substantial gene flow among groups with small populations. It is likely that the relatively high values for Fsc in comparison with global Fst are a reflection of this

phenomenon. Consequently, almost all of the overall variance is attributable to differences within geographic regions (Fsc), rather than between them (Fct). A similar pattern, albeit resulting in significant differences between regions, was observed for mtDNA Fst, where the majority of the overall variance was attributable to differences within geographic regions, but a significant minority of the overall variance was also attributable to differences between geographic regions. This highlights differences in the overall scale of geographic structure of haplotypes observed in the two sex-specific systems, with structure observed over much finer geographic scales for NRY haplotypes than that observed for mtDNA haplotypes.

Table 3.9 AMOVA results, performed using samples assigned to the ethnic groups, grouped by geographic region (N,S,E,W)

	Fsc	Fct	Global Fst (ungrouped data)
mtDNA Fst	0.009 (<0.001)	0.003 (0.003)	0.010 (<0.001)
mtDNA K2P	0.014 (<0.001)	0.020 (<0.001)	0.021 (<0.001)
NRY UEP Fst	0.092 (<0.001)	0.109 (0.002)	0.130 (<0.001)
NRY UEP-MS Fst	0.057 (<0.001)	**0.004 (0.234)**	0.058 (<0.001)
NRY MS Rst	0.097 (<0.001)	0.079 (0.010)	0.124 (<0.001)

3.3 Are Omotic and Cushitic speakers more similar to each other than either is to Nilo-Saharan or Semitic speakers?

The modal first language linguistic group was Omotic, which was spoken by 37.1% of the sample donors (2,138 of 5,756, Table 3.10). Cushitic languages were spoken by 35.3% of sample donors, and Semitic and Nilo-Sahran languages spoken by 21.7% and 5.9% of sample donors respectively. A single Amharan sample donor spoke English (Indo-European) as his first language, and was consequently excluded from further analysis.

Table 3.10 Distribution of samples from ethnic groups amongst the first language linguistic group of the sample donor.

Ethnic group	Cushitic	Nilo-Saharan	Indo-European	Omotic	Semitic	Grand Total
AA				102	15	117
AF	111				1	112
AG	152				117	269
AL	102				8	110
AM	2		1		393	396
AN		108				108
BE				122	2	124
BK				110	3	113
BN				126	1	127
BR	119					119
BU	115			11		126
DI				116	16	132
DR	79			1	28	108
DS	102				3	105
DW				111	6	117
DZ	1			100	3	104
GB	113					113
GE	122					122
GF				47	64	111
GG				109		109
GM				70	139	209
GN	2			91	20	113
GR	2				150	152
GW	116				1	117
HD	106				21	127
HM				107	5	112
KF	3			106	11	120
KM	105			1	11	117
KN				107		107
KR	18			74	16	108
KS	94			2	24	120
MG		115				115
MH	130					130
ML				115	4	119
NR		118				118
OR	108				41	149
SC	2			118	5	125
SD	110				16	126
SK				112	1	113
SM	100				8	108
TG	1				65	66
TY	113			1		114
WL	1			72	37	110
YM				107	1	108
ZS				100	11	111
Grand Total	2029	341	1	2138	1247	5756

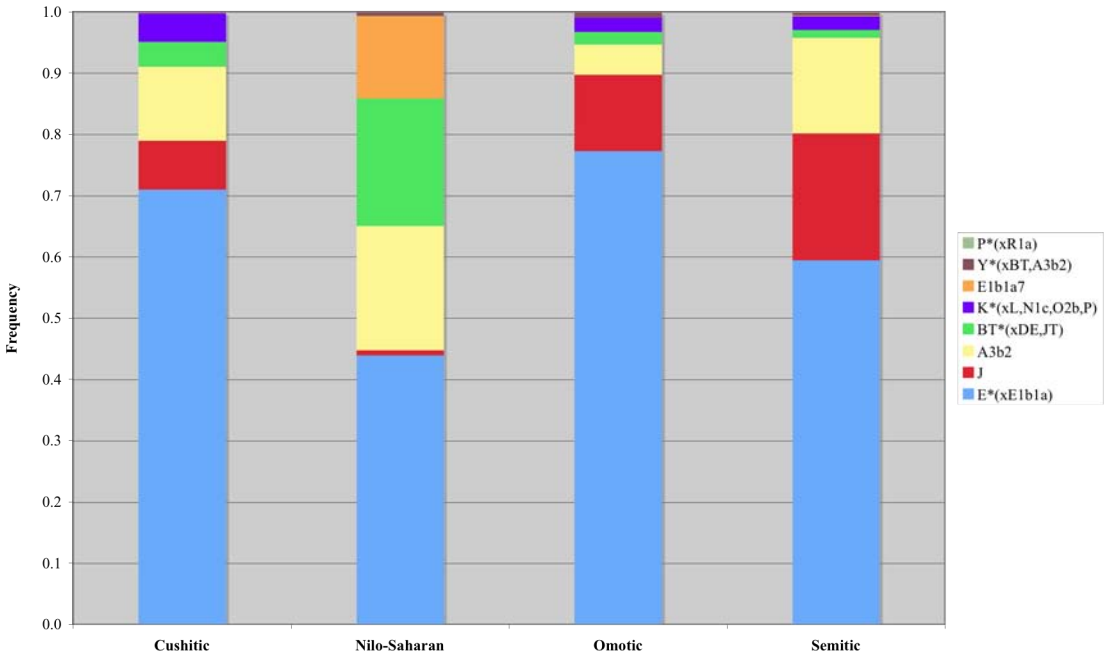
3.3.1 NRY diversity

The greatest gene diversity (Table 3.11) at NRY haplogroup level was found in Nilo-Saharan speakers, with an *h* value of 0.706, whereas the lowest *h* was found in Omotic speakers (*h*=0.383) (Z test p<0.01). The mean and median haplogroup level gene diversity was 0.535 and 0.525 respectively. The highest frequency of the modal Ethiopian haplogroup, E*(xE1b1a), was found in Omotic speakers (77.3%, Figure 3.43), whereas the lowest frequency was found in Nilo-Saharan speakers (44.0%). Additionally, Nilo-Saharan speakers had the highest frequencies of haplogroups A3b2 and BT*(xDE,JT) (20.2% and 20.8% respectively), as well as haplogroup E1b1a7 (13.5%), a haplogroup that did not occur at substantial frequencies in any other linguistic group. At the haplotype level, the highest gene diversity was found in Cushitic speakers (*h*=0.977), whereas the lowest was found in Nilo-Saharan speakers (*h*=0.945) (Z test p<0.01), with mean and median values of 0.966 and 0.971 respectively. The lowest mean microsatellite repeat length variance also occurred in Nilo-Saharan speakers, with an MSV of 0.721, whereas the highest was found for Semitic speakers (1.134). Mean MSV was 0.887, and the median MSV was 0.845.

Table 3.11 Summary of the genetic diversity found within linguistic groups

Linguistic group	NRY haplogroup *h*	s.d ±	NRY haplotype *h*	s.d ±	NRY MS MSV	mtDNA haplotype *h*	s.d ±	mtDNA haplotype π	s.d ±
Cushitic	0.471	0.011	0.977	0.003	0.926	0.994	0.002	0.0248	0.0126
Nilo-	0.706	0.025	0.945	0.012	0.721	0.990	0.005	0.0262	0.0133
Omotic	0.383	0.011	0.967	0.004	0.765	0.992	0.002	0.0240	0.0122
Semitic	0.579	0.014	0.974	0.005	1.134	0.994	0.002	0.0241	0.0123

Figure 3.43 Frequencies of NRY haplogroups found in linguistic groups



3.3.2 mtDNA diversity

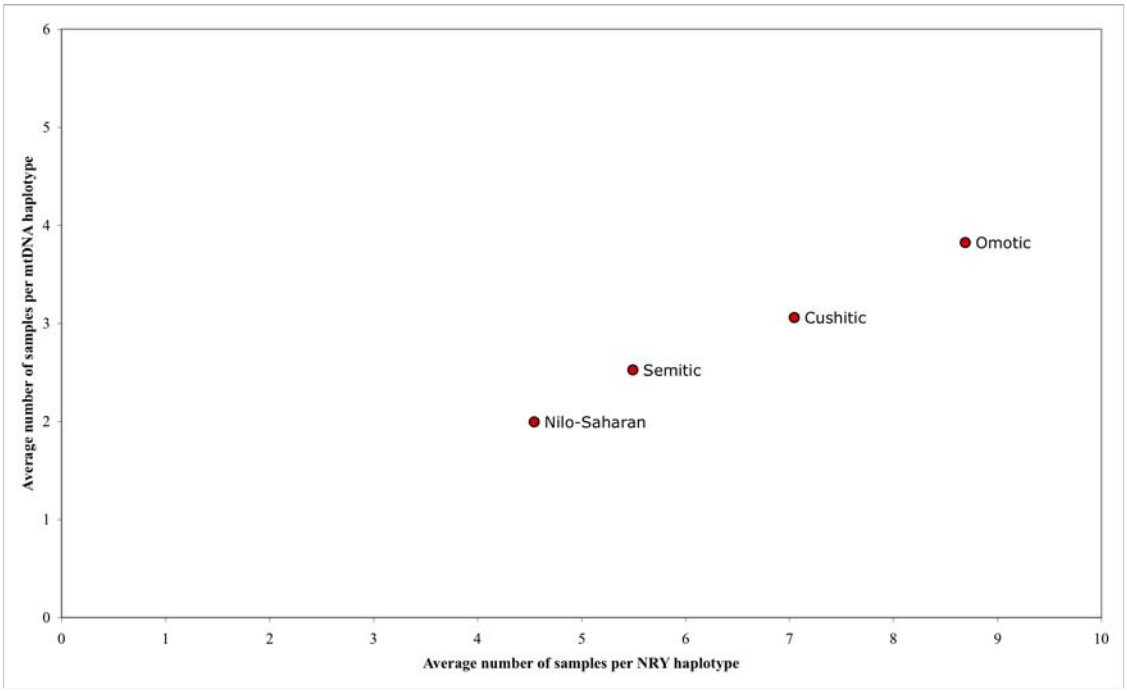
Haplotype level gene diversity ranged from 0.994 in both the Cushitic and Semitic speakers, to 0.990 in Nilo-Saharan speakers (although not significantly different, Z test p=0.42). The mean haplotype gene diversity was 0.992, and the median value 0.993. Nilo-Saharan speakers had the highest nucleotide diversity (0.0262), whereas Omotic speakers were found to have the lowest (0.0241) (although due to the large standard deviation, these value were not significantly different, Z test p=0.87). Mean and median nucleotide diversity values were 0.0248 and 0.0244.

3.3.3 Comparison of NRY and mtDNA diversity

A significant linear correlation was observed between the numbers of samples per NRY haplotype and the numbers of samples per mtDNA HVS1 haplotype in linguistic groups (r^2=0.9941, p=0.0030, Figure 3.44). Cushitic speakers showed the highest haplotype level gene diversity for both the NRY and mtDNA data, but had intermediate diversity values for all other datasets and metrics (Table 3.11). Nilo-Saharan speakers had both the highest NRY UEP level gene diversity, and the highest mtDNA nucleotide diversity. However, Nilo-Saharan speakers also showed the lowest gene diversity for NRY and mtDNA haplotypes, as well as the lowest MSV. Omotic speakers exhibited both the lowest UEP haplogroup level gene diversity as well as the lowest mtDNA nucleotide

diversity, with all other diversity measures showing intermediate values. Semitic speakers were notable for exhibiting the highest MSV and the second highest mtDNA haplotype gene diversity, although all other diversity measures had intermediate values.

Figure 3.44 Plot of the average number of samples per haplotype per linguistic group for NRY and mtDNA



3.3.4 Exact Tests of Population Differentiation between linguistic groups

ETPD performed using NRY UEP haplogroup frequencies, NRY UEP-MS haplotype frequencies and mtDNA HVS1 haplotype frequencies all showed that all linguistic groups were significantly differentiated from each other ($p<0.001$ between all pairs of linguistic groups, Table 3.12-Table 3.14).

Table 3.12 Matrix of ETPD p values based on NRY UEP haplogroup frequencies in linguistic groups

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*			
Nilo-Saharan	<0.001	*		
Omotic	<0.001	<0.001	*	
Semitic	<0.001	<0.001	<0.001	*

Table 3.13 Matrix of ETPD p values based on NRY UEP-MS haplotype frequencies in linguistic groups

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*			
Nilo-Saharan	<0.001	*		
Omotic	<0.001	<0.001	*	
Semitic	<0.001	<0.001	<0.001	*

Table 3.14 Matrix of ETPD p values based on mtDNA HVS1 haplotype frequencies in linguistic groups

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*			
Nilo-Saharan	<0.001	*		
Omotic	<0.001	<0.001	*	
Semitic	<0.001	<0.001	<0.001	*

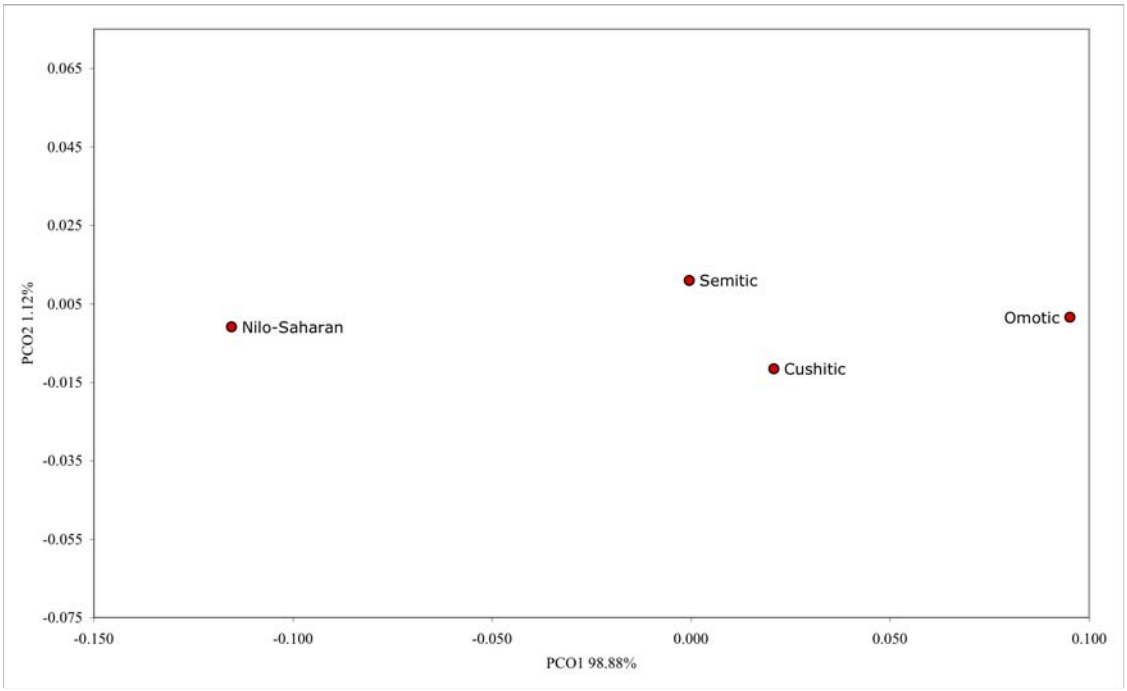
3.3.5 Genetic distances between linguistic groups

Pairwise Fst distances between linguistic groups using frequencies of NRY UEP haplogroups showed the largest difference to be between Nilo-Saharan speakers and Omotic speakers (Fst=0.2102, Table 3.15). The smallest difference was between Cushitic and Omotic speakers (Fst=0.0136), with Semitic speakers appearing closer to Cushitic speakers than to Omotic speakers (Figure 3.45).

Table 3.15 Pairwise Fst distances (lower diagonal) with associated p values (upper diagonal) between linguistic groups using NRY UEP haplogroup frequencies

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*	<0.001	<0.001	<0.001
Nilo-Saharan	0.1222	*	<0.001	<0.001
Omotic	0.0136	0.2102	*	<0.001
Semitic	0.0305	0.0943	0.0533	*

Figure 3.45 PCO of pairwise Fst distances between linguistic groups using UEP haplogroup frequencies

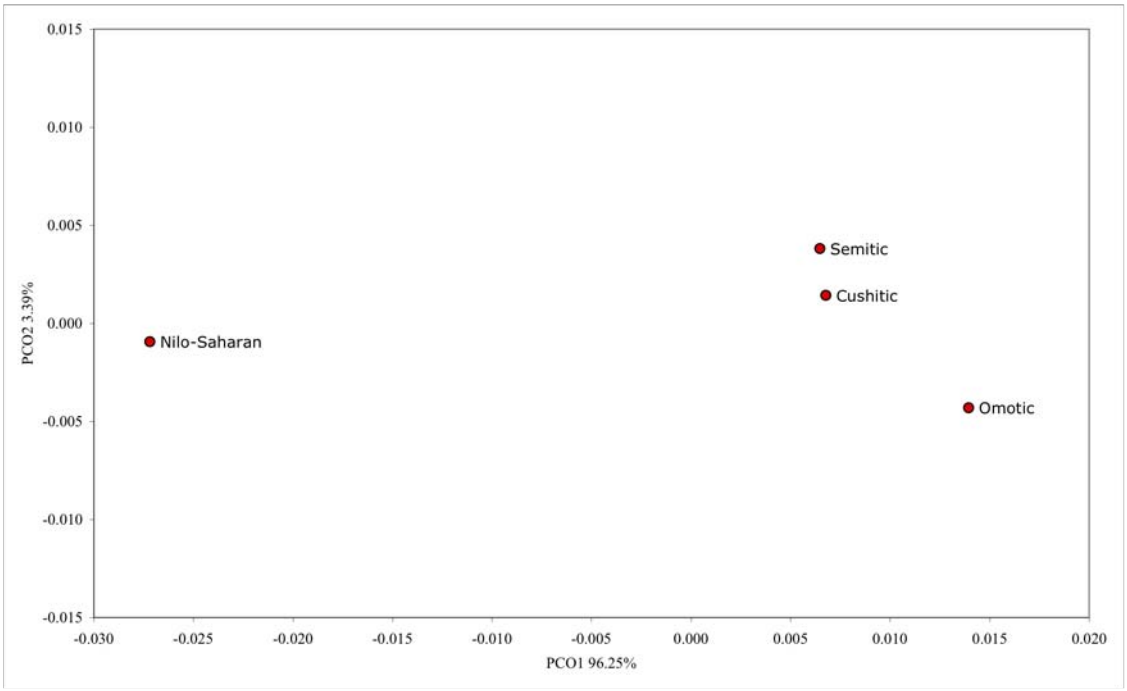


Using UEP-MS haplotype frequencies, pairwise Fst distances were largest between Nilo-Saharan speakers and Omotic speakers (Fst=0.0413, Table 3.16). The smallest distance was between Cushitic and Semitic speakers (Fst=0.0036). Cushitic and Semitic speakers are almost equally distant from Nilo-Saharan speakers (Fst=0.0341 and 0.0340 respectively), as well as similarly distant from Omotic speakers (Fst=0.0094 and 0.0110 respectively).

Table 3.16 Pairwise Fst distances (lower diagonal) with associated p values (upper diagonal) between linguistic groups using NRY UEP-MS haplotype frequencies

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*	<0.001	<0.001	<0.001
Nilo-Saharan	0.0341	*	<0.001	<0.001
Omotic	0.0094	0.0413	*	<0.001
Semitic	0.0036	0.0340	0.0110	*

Figure 3.46 PCO of pairwise Fst distances between linguistic groups using UEP-MS haplotype frequencies

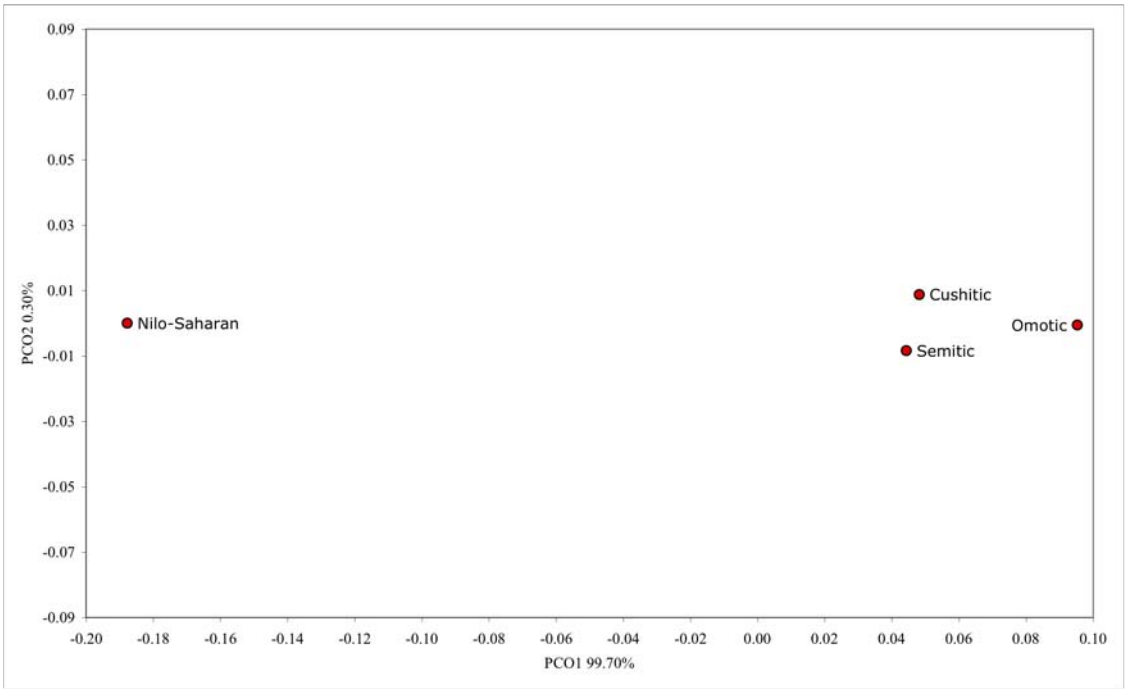


Similar to pairwise Fst distances using NRY data, pairwise Rst distances using MS frequencies showed the greatest distance to be between Nilo-Saharan speakers and Omotic speakers (Rst=0.2821, Table 3.17). The smallest Rst distance was between Cushitic and Semitic speakers (Rst=0.0173), with both of these groups similarly distant from Nilo-Saharan speakers (Fst =0.2350 and 0.2309 respectively). Omotic speakers were slightly closer to Cushitic than to Semitic speakers (Fst=0.0196 and 0.0222 respectively), although all three appear to cluster relative to Nilo-Saharan speakers (Figure 3.47).

Table 3.17 Pairwise Rst distances (lower diagonal) with associated p values (upper diagonal) between linguistic groups using NRY MS frequencies

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*	<0.001	<0.001	<0.001
Nilo-Saharan	0.2350	*	<0.001	<0.001
Omotic	0.0196	0.2821	*	<0.001
Semitic	0.0173	0.2309	0.0222	*

Figure 3.47 PCO of pairwise Rst distances between linguistic groups using MS frequencies

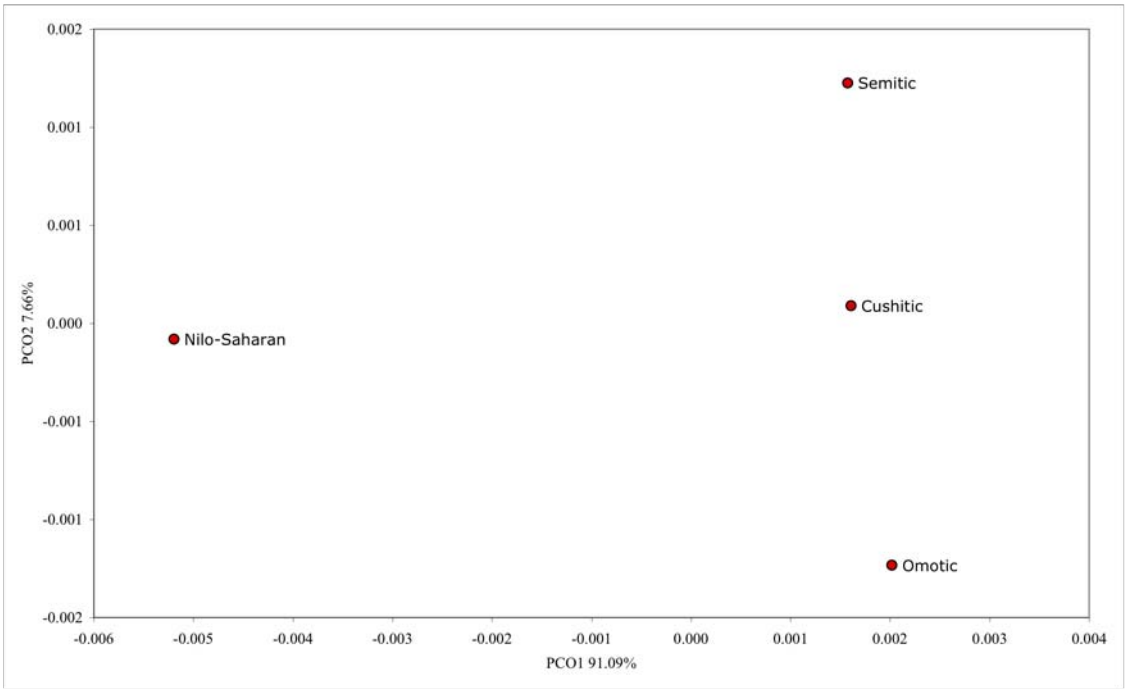


Pairwise Fst distances using mtDNA HVS1 haplotype frequencies showed the greatest distance to be between Nilo-Saharan speakers and Omotic speakers (Fst=0.0073, Table 3.18). The distance between Nilo-Saharan speakers and Cushitic and Semitic speakers were similar (Fst=0.0068 and 0.0069 respectively) and not very different to the Nilo-Saharan-Omotic distance. The smallest distance was between Cushitic and Semitic speakers (Fst=0.0014), with Omotic speakers appearing closer to Cushitic than Semitic groups (Fst=0.0016 and 0.0025 respectively).

Table 3.18 Pairwise Fst distances (lower diagonal) with associated p values (upper diagonal) between linguistic groups using mtDNA HVS1 haplotype frequencies

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*	<0.001	<0.001	<0.001
Nilo-Saharan	0.0068	*	<0.001	<0.001
Omotic	0.0016	0.0073	*	<0.001
Semitic	0.0014	0.0069	0.0025	*

Figure 3.48 PCO of pairwise Fst distances between linguistic groups using mtDNA HVS1 haplotype frequencies

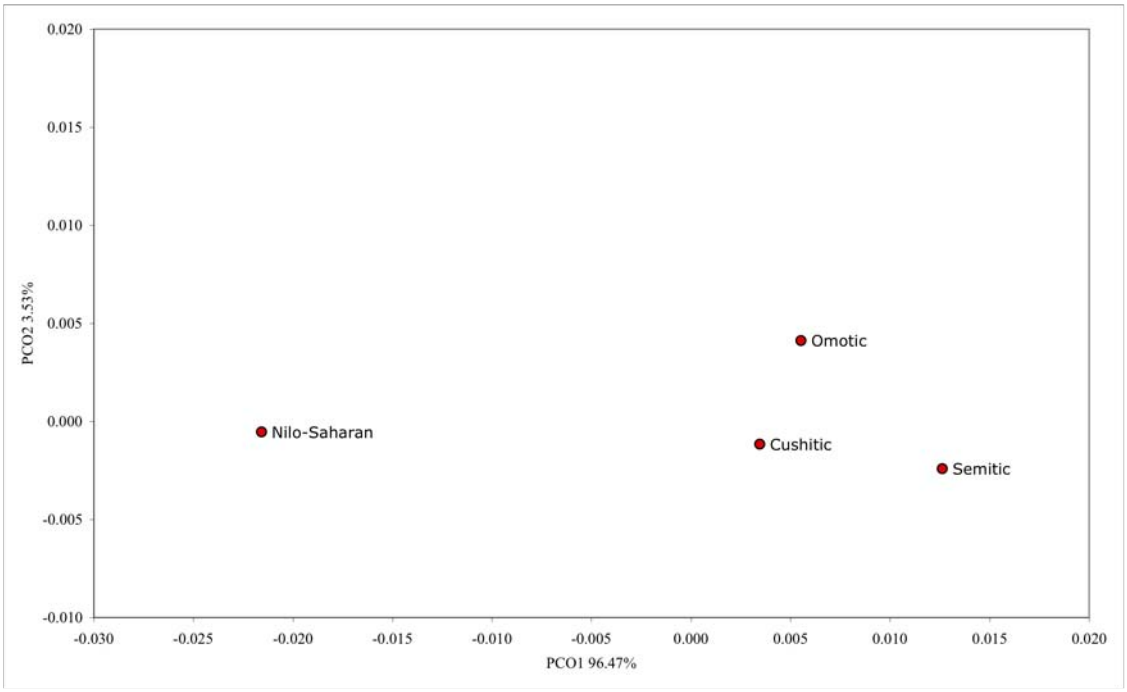


The pattern observed for pairwise K2P distances was different from those for Fst distances using mtDNA HVS1 haplotypes between linguistic groups. The largest distance was between Nilo-Saharan and Semitic speakers (K2P=0.0342, Table 3.19). The smallest distance was between Cushitic and Omotic speakers (K2P=0.0021), with Semitic speakers appearing to be closer to Cushitic than Omotic speakers (K2P=0.0052 and 0.0093 respectively). Nilo-Saharan speakers were closer to Cushitic speakers than the other groups (K2P=0.0244).

Table 3.19 Pairwise K2P distances (lower diagonal) with associated p values (upper diagonal) between linguistic groups using mtDNA HVS1 haplotypes

	Cushitic	Nilo-Saharan	Omotic	Semitic
Cushitic	*	<0.001	<0.001	<0.001
Nilo-Saharan	0.0244	*	<0.001	<0.001
Omotic	0.0021	0.0275	*	<0.001
Semitic	0.0052	0.0342	0.0093	*

Figure 3.49 PCO of pairwise K2P distances between linguistic groups using mtDNA HVS1 haplotypes



Chapter 4: Genetic and social patterns of ethnicity displayed by and associated with sex specific genetic systems

4.1 How different is the ethnicity of the sampled generation to that of previous generations?

Table 4.1 shows the proportion of donors with a declared ethnicity identical to, respectively, their fathers, mothers, paternal grandfathers and maternal grandmothers (see Supplementary Table Ethnic). The most ethnically uniform groups, with all generations sharing the same ethnicity, were the Nuer (NR) and Sheko (SK). The group displaying the greatest difference was the Gamo (GM), with only 37.8% of Gamo samples having the same declared ethnicity as all of their parents, paternal grandfather and maternal grandmother. For the entire dataset, it is significantly more likely that the donor's ethnicity will match that of their father than that of their mother (mean values for the proportion of sample donors with ethnicities that match either their father's or mother's being 0.941 and 0.901 respectively (2 tailed t-test $p<0.0001$)). The largest difference in donor's ethnicities matching that of their father rather than that of their mother is seen in the Kore (KR), with all sample donor's ethnicities identical to those of their father's, but only 81.5% matching the ethnicity of their mother. The largest difference in the donor's ethnicity matching their mother's rather than their father's ethnicity occurred in the Dizi (DI), with 95.5% of sample donor's matching their mother's ethnicity, compared with 86.4% of donor's matching the ethnicity of their father. Interestingly, it is more common for the ethnicity of the donor's father to match that of the donor's paternal grandfather, than for the ethnicity of the donor to match that of the donor's father (mean match values of 0.976 and 0.941 respectively, $p<0.0001$). Likewise, it is more likely for the donor's mother's ethnicity to be the same as that of the donor's maternal grandmother, than for the donor's ethnicity to match that of their mother (mean match values of 0.967 and 0.901 respectively, $p<0.0001$). This may indicate that the label of self-declared ethnic identity is more fluid in the current generation than in the generation preceding it. Because all donors were male, ethnicity match values could be different for females of the current generation. Further evidence for the plasticity of ethnicity is the existence of donors with ethnicities that do not match

that of either of their parents. Donors whose ethnicities did not match those of either parent were found at low frequencies (ranging from 0.8% to 11.0% of samples) in 30 of the 45 ethnic groups, with the highest frequency of occurrence in the Gamo (GM, 11.0% of samples). There was clear evidence of recent admixture as measured by the frequency of inter-ethnic marriage in the dataset, with a mean value of 12.6% of donors with parents drawn from two different ethnic groups. The highest frequency of this was seen in the Gamo (GM), with 48.3% of donors having parents of two different ethnic groups, whereas the Nuer, Sheko and Gewada (NR, SK and GW respectively) donors were all sons of marriages between members of the same ethnic group.

Table 4.1 Proportion of ethnicity matches between sample donors, donor's parent's, paternal grandfathers and maternal grandmothers, in ethnic groups of the sample donor.

Ethnic group	D=F	F=FF	D=M	M=MM	F=M	D=F=M	D=F=M=FF=MM	D≠ForM
AA	0.940	0.966	0.957	0.966	0.906	0.906	0.897	0.009
AF	1.000	0.982	0.938	0.991	0.938	0.938	0.911	0.000
AG	0.948	0.963	0.892	0.941	0.870	0.855	0.810	0.015
AL	0.964	0.991	0.827	0.918	0.827	0.809	0.755	0.018
AM	0.932	0.962	0.934	0.962	0.907	0.889	0.851	0.023
AN	0.991	1.000	1.000	1.000	0.991	0.991	0.991	0.000
BE	1.000	0.992	0.960	0.984	0.960	0.960	0.944	0.000
BK	0.938	0.991	0.956	0.973	0.894	0.894	0.876	0.000
BN	1.000	0.992	0.921	1.000	0.921	0.921	0.921	0.000
BR	0.992	1.000	0.924	0.992	0.916	0.916	0.908	0.000
BU	0.921	0.984	0.905	0.960	0.881	0.857	0.817	0.032
DI	0.864	0.970	0.955	0.977	0.833	0.826	0.818	0.008
DR	0.917	0.917	0.815	0.917	0.778	0.759	0.685	0.028
DS	0.981	1.000	0.914	0.990	0.933	0.914	0.905	0.019
DW	0.974	0.991	0.932	0.974	0.906	0.906	0.872	0.000
DZ	0.971	0.981	0.962	1.000	0.942	0.942	0.933	0.010
GB	0.982	0.956	0.956	0.956	0.947	0.947	0.876	0.009
GE	0.984	1.000	1.000	1.000	0.984	0.984	0.984	0.000
GF	0.856	0.928	0.712	0.946	0.640	0.604	0.559	0.036
GG	0.917	1.000	0.908	0.991	0.899	0.890	0.890	0.064
GM	0.689	0.890	0.670	0.856	0.517	0.469	0.378	0.110
GN	0.947	0.982	0.894	0.991	0.841	0.841	0.832	0.000
GR	0.901	0.947	0.862	0.921	0.829	0.803	0.743	0.039
GW	0.991	1.000	0.991	0.991	1.000	0.991	0.983	0.009
HD	0.921	0.976	0.803	0.913	0.772	0.748	0.677	0.024
HM	0.991	1.000	0.946	0.973	0.938	0.938	0.938	0.000
KF	0.892	0.975	0.950	0.992	0.892	0.867	0.858	0.025
KM	0.966	0.966	0.821	0.906	0.795	0.795	0.709	0.009
KN	0.935	0.991	0.963	0.991	0.935	0.916	0.897	0.019
KR	1.000	1.000	0.815	0.991	0.815	0.815	0.815	0.000
KS	0.900	0.983	0.867	0.983	0.825	0.817	0.817	0.050
MG	0.991	0.991	0.983	1.000	0.974	0.974	0.965	0.000
MH	0.969	0.962	0.931	0.977	0.931	0.923	0.885	0.023
ML	0.950	0.958	0.849	0.975	0.899	0.849	0.815	0.050
NR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
OR	0.933	0.980	0.866	0.919	0.826	0.812	0.765	0.013
SC	0.952	0.976	0.888	0.992	0.856	0.856	0.848	0.016
SD	0.952	0.984	0.913	0.992	0.905	0.897	0.897	0.032
SK	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
SM	0.926	0.991	0.898	0.981	0.861	0.843	0.833	0.019
TG	0.818	0.970	0.682	0.939	0.591	0.545	0.530	0.045
TY	0.982	1.000	0.974	1.000	0.974	0.965	0.965	0.009
WL	0.900	0.955	0.773	0.955	0.727	0.709	0.664	0.036
YM	0.981	1.000	0.991	0.981	0.972	0.972	0.963	0.000
ZS	0.955	0.991	0.910	0.991	0.901	0.883	0.874	0.018
Ethiopia mean	0.941	0.976	0.901	0.967	0.874	0.862	0.835	0.020

D = Donor, F = Father, M = Mother, FF = Paternal grandfather, MM = Maternal grandmother

Figure 4.1 Proportion of ethnicity matches between the sample donor and both the donor's parents in ethnic groups of the sample donor (ordered by increasing frequency of the sample donor's ethnicity identical to both parents)

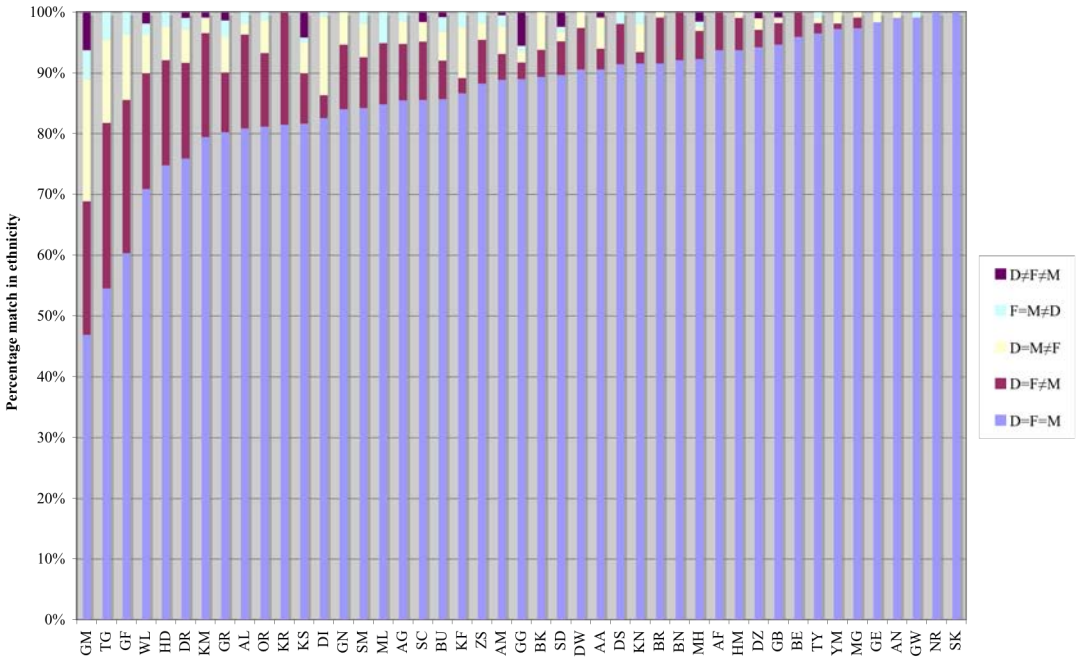


Table 4.2 shows the diversity values (Nei's h) in ethnic groups for the donor, donor's father, mother, paternal grandfather and maternal grandmother, calculated using the frequency of ethnic groups as shown in Supplementary Table Ethnic. The mean diversity for the donor's father was 0.103, for the donor's mother was 0.176, for the paternal grandfather was 0.121, for the maternal grandmother was 0.194, and the mean diversity across all ethnic groups and generations was 0.148. The largest diversity in ethnicity values across all generations for an ethnic group occurred in the Gamo (GM), with values 0.504, 0.528, 0.580 and 0.570 for the donor's father, mother, paternal grandfather and maternal grandmother respectively. The lowest diversity in ethnicity occurred in the Nuer (NR) and Sheko (SK), with parents, paternal grandfather and maternal grandmother all having the same ethnicity as the sample donor.

From Supplementary Table Ethnic, it is clear that overall the most common ethnicity of the donor's parents and grandparents, if not the same as the sample donor, was Amhara. Amhara represents 6.9% (n=396) of ethnicities of the samples donor, but overall represents 8.1% of the ethnicities for donor's fathers, 9.0% of the ethnicities for donor's mothers, 8.3% of the ethnicities of the donor's paternal grandfather and 9.2% of the

ethnicities of the donor's maternal grandmother. Oromo as an ethnicity was also far more common for the donor's parents and grandparents, than for sample donor's, representing 2.6% (n=149) of sample donor's ethnicities, but 3.7% of the ethnicities for both the donor's parents, 3.8% for donor's paternal grandfathers and 3.9% for donor's maternal grandmothers.

Table 4.2 Diversity values for the ethnicity of the donor's parents, paternal grandfather and maternal grandmother in ethnic groups of the sample donor

Ethnic group	Father	s.d ±	Mother	s.d ±	Paternal grandfather	s.d ±	Maternal grandmother	s.d ±
AA	0.116	0.030	0.084	0.026	0.100	0.028	0.100	0.028
AF	0.000	0.000	0.120	0.031	0.035	0.017	0.137	0.032
AG	0.100	0.018	0.197	0.024	0.141	0.021	0.227	0.026
AL	0.071	0.025	0.311	0.044	0.089	0.027	0.383	0.046
AM	0.130	0.017	0.126	0.017	0.171	0.019	0.154	0.018
AN	0.019	0.013	0.000	0.000	0.019	0.013	0.000	0.000
BE	0.000	0.000	0.079	0.024	0.016	0.011	0.109	0.028
BK	0.119	0.031	0.087	0.026	0.136	0.032	0.120	0.031
BN	0.000	0.000	0.150	0.032	0.016	0.011	0.150	0.032
BR	0.017	0.012	0.145	0.032	0.017	0.012	0.160	0.034
BU	0.152	0.032	0.181	0.034	0.181	0.034	0.249	0.039
DI	0.250	0.038	0.088	0.025	0.199	0.035	0.074	0.023
DR	0.159	0.035	0.334	0.045	0.210	0.039	0.420	0.047
DS	0.038	0.019	0.163	0.036	0.038	0.019	0.180	0.038
DW	0.051	0.020	0.132	0.031	0.067	0.023	0.179	0.035
DZ	0.057	0.023	0.076	0.026	0.076	0.026	0.076	0.026
GB	0.035	0.017	0.086	0.026	0.087	0.026	0.168	0.035
GE	0.033	0.016	0.000	0.000	0.033	0.016	0.000	0.000
GF	0.266	0.042	0.470	0.047	0.326	0.044	0.426	0.047
GG	0.158	0.035	0.175	0.036	0.158	0.035	0.175	0.036
GM	0.504	0.035	0.528	0.035	0.580	0.034	0.570	0.034
GN	0.104	0.029	0.201	0.038	0.137	0.032	0.201	0.038
GR	0.185	0.032	0.252	0.035	0.219	0.034	0.285	0.037
GW	0.017	0.012	0.017	0.012	0.017	0.012	0.034	0.017
HD	0.149	0.032	0.345	0.042	0.163	0.033	0.433	0.044
HM	0.018	0.013	0.104	0.029	0.018	0.013	0.104	0.029
KF	0.202	0.037	0.097	0.027	0.203	0.037	0.081	0.025
KM	0.067	0.023	0.314	0.043	0.132	0.031	0.353	0.044
KN	0.126	0.032	0.073	0.025	0.143	0.034	0.091	0.028
KR	0.000	0.000	0.332	0.045	0.000	0.000	0.317	0.045
KS	0.190	0.036	0.248	0.039	0.205	0.037	0.247	0.039
MG	0.017	0.012	0.034	0.017	0.035	0.017	0.034	0.017
MH	0.061	0.021	0.134	0.030	0.133	0.030	0.175	0.033
ML	0.098	0.027	0.273	0.041	0.145	0.032	0.287	0.041
NR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
OR	0.127	0.027	0.240	0.035	0.161	0.030	0.267	0.036
SC	0.094	0.026	0.206	0.036	0.078	0.024	0.219	0.037
SD	0.093	0.026	0.166	0.033	0.093	0.026	0.166	0.033
SK	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SM	0.142	0.034	0.192	0.038	0.124	0.032	0.192	0.038
TG	0.322	0.058	0.513	0.062	0.324	0.058	0.493	0.062
TY	0.035	0.017	0.052	0.021	0.035	0.017	0.052	0.021
WL	0.187	0.037	0.396	0.047	0.251	0.041	0.422	0.047
YM	0.037	0.018	0.019	0.013	0.037	0.018	0.037	0.018
ZS	0.088	0.027	0.171	0.036	0.088	0.027	0.187	0.037

4.2 How does the first language spoken by the sampled generation compare with previous generations?

Table 4.3 shows the proportional matches between the first language spoken by the sample donor and the donor's parents, paternal grandfather and maternal grandmother (see Supplementary Table Language). The most uniform ethnic group with regard to inter-generation linguistic diversity was the Nuer (NR), with the sample donor, donor's father, mother, paternal grandfather and maternal grandmother all speaking the same first language. The most diverse ethnic group with regard to linguistic ancestry was the Gamo (GM), with 29.2% of Gamo donors speaking the same first language as both of their parents, and paternal grandfather and maternal grandmother. The highest incidences where the donor's first language matches that of their father rather than that of their mother occurred in the Alaba (AL), with match values of 0.173 and 0.027 respectively. The highest incidence where the donor's first language matches that of their mother rather than that of their father occurred in the Tigray (TG), with match values of 0.258 and 0.076 respectively. Overall however, there was no significant difference in whether the donor's first language matched that of one parent rather than the other (2 tailed t-test $p=0.306$). It was significantly more likely for the first language spoken by the donor's father to be that of the donor's paternal grandfather, than for the language of the donor to be the same as that of their father (mean match values of 0.963 and 0.855 respectively, $p<0.0001$). Similarly, it was significantly more likely for the donor's mother's first language to be that of the maternal grandmother, than for the donor's first language to be the same as that of their mother (mean match values of 0.958 and 0.846 respectively, $p<0.0001$). On average, 86.5% of sample donor's had parents that spoke the same first language, which ranged from 100% in the Nuer and Sheko (NR and SK respectively), to 54.1% in the Gamo (GM). The mean frequency of sample donor's for whom their first language is not spoken as a first language by either parent was 10.0%, and was observed at highest frequency in the Gofa (GF, 56.8%), but the phenomenon was seen in 42 of the 45 ethnic groups.

Table 4.3 Proportion of first language matches between sample donors and donor's parents, paternal grandfathers and maternal grandmothers in ethnic groups of the sample donor.

Ethnic group	D=F	F=FF	D=M	M=MM	F=M	D=F=M	D=F=M=FF=MM	D≠ForM
AA	0.923	0.966	0.906	0.974	0.923	0.897	0.872	0.068
AF	0.982	0.991	0.929	0.991	0.929	0.920	0.902	0.009
AG	0.784	0.862	0.695	0.903	0.751	0.628	0.546	0.149
AL	0.891	0.991	0.745	0.918	0.800	0.718	0.691	0.082
AM	0.927	0.972	0.942	0.960	0.909	0.891	0.851	0.023
AN	0.991	0.991	1.000	1.000	0.991	0.991	0.981	0.000
BE	0.992	0.992	0.960	0.992	0.960	0.960	0.952	0.008
BK	0.920	0.982	0.947	0.982	0.894	0.885	0.867	0.018
BN	0.984	0.992	0.921	1.000	0.921	0.913	0.913	0.008
BR	0.706	0.983	0.697	0.992	0.899	0.672	0.672	0.269
BU	0.897	0.992	0.857	0.968	0.881	0.825	0.810	0.071
DI	0.841	0.970	0.879	0.970	0.826	0.788	0.780	0.068
DR	0.676	0.889	0.657	0.898	0.778	0.620	0.565	0.287
DS	0.981	0.990	0.933	0.990	0.943	0.933	0.914	0.019
DW	0.915	0.991	0.923	0.966	0.906	0.906	0.863	0.068
DZ	0.923	0.971	0.942	1.000	0.933	0.913	0.904	0.048
GB	0.938	0.956	0.912	0.956	0.929	0.894	0.850	0.044
GE	0.975	1.000	0.984	0.992	0.975	0.967	0.967	0.008
GF	0.378	0.919	0.378	0.955	0.631	0.324	0.306	0.568
GG	0.917	1.000	0.908	0.991	0.899	0.890	0.890	0.064
GM	0.545	0.847	0.598	0.818	0.541	0.402	0.292	0.258
GN	0.717	0.982	0.735	0.991	0.841	0.690	0.681	0.239
GR	0.671	0.875	0.638	0.836	0.770	0.559	0.434	0.250
GW	0.966	0.983	0.974	0.983	0.991	0.966	0.957	0.026
HD	0.795	0.961	0.709	0.913	0.764	0.654	0.598	0.150
HM	0.946	0.991	0.946	0.982	0.955	0.929	0.929	0.036
KF	0.875	0.967	0.867	0.992	0.883	0.817	0.808	0.075
KM	0.872	0.957	0.769	0.915	0.795	0.735	0.658	0.094
KN	0.935	0.991	0.963	0.991	0.935	0.916	0.897	0.019
KR	0.667	0.991	0.667	0.972	0.806	0.620	0.620	0.287
KS	0.758	0.983	0.775	0.967	0.817	0.733	0.733	0.200
MG	0.991	0.991	0.983	1.000	0.974	0.974	0.965	0.000
MH	0.938	0.969	0.908	0.977	0.923	0.900	0.862	0.054
ML	0.916	0.983	0.840	0.966	0.882	0.832	0.815	0.076
NR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
OR	0.725	0.966	0.826	0.933	0.805	0.678	0.631	0.128
SC	0.896	0.976	0.912	0.984	0.856	0.856	0.848	0.048
SD	0.889	0.976	0.921	0.984	0.913	0.881	0.865	0.071
SK	0.991	1.000	0.991	1.000	1.000	0.991	0.991	0.009
SM	0.907	0.972	0.898	0.972	0.861	0.843	0.824	0.037
TG	0.485	0.939	0.667	0.894	0.561	0.409	0.379	0.258
TY	0.991	1.000	0.965	1.000	0.974	0.965	0.965	0.009
WL	0.718	0.900	0.700	0.900	0.773	0.618	0.536	0.200
YM	0.963	0.981	0.991	0.981	0.963	0.963	0.954	0.009
ZS	0.856	0.973	0.838	0.973	0.901	0.820	0.802	0.126
Ethiopia mean	0.855	0.963	0.846	0.958	0.865	0.801	0.771	0.100

D = Donor, F = Father, M = Mother, FF = Paternal grandfather, MM = Maternal grandmother

Figure 4.2 Proportion of first language matches between the sample donor and both the donor's parents in ethnic groups of the sample donor (ordered by increasing frequency of the sample donor's first language identical to both parents)

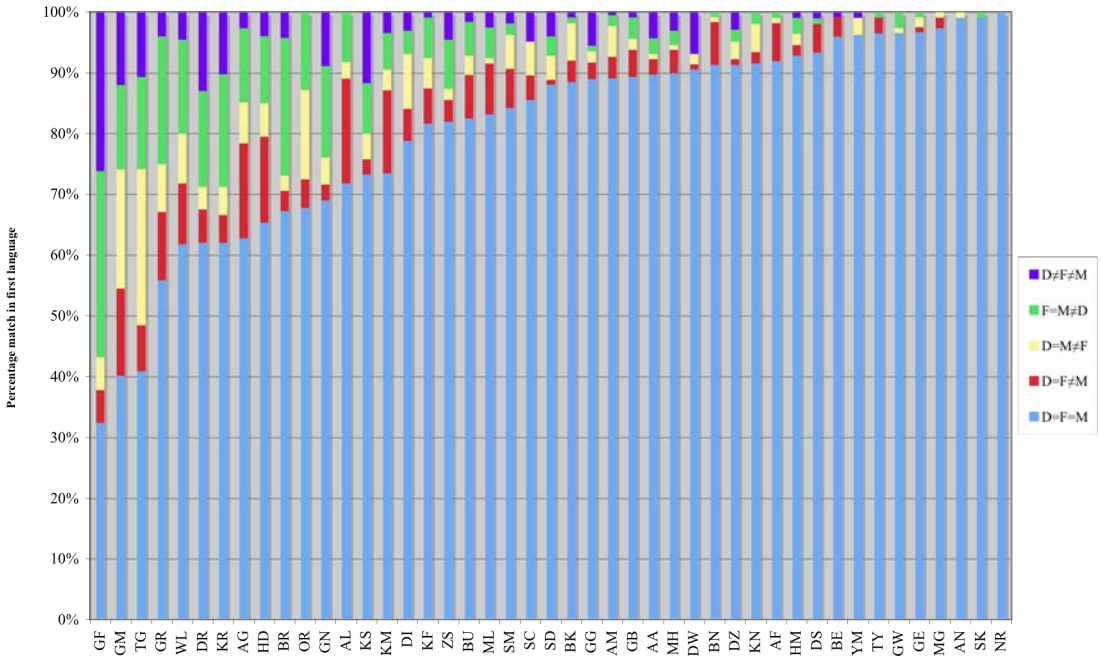


Table 4.4 shows the diversity values in ethnic groups for the sample donor, donor's father, mother, paternal grandfather and maternal grandmother, calculated using the frequencies of first languages as shown in Supplementary Table Language. When comparing the frequency of first languages in different generations Supplementary Table Language, the most noticeable feature appears to be that the frequency of Amharic (AM) as a first language is far higher in the donor's generation (19.8%), than in the father's (11.0%), mother's (11.9%), paternal grandfather's (9.9%) or maternal grandmother's (10.8%) generation. Correspondingly, the frequencies of other first languages (i.e. excluding each group's modal tongue) are lower in the donor's generation compared with other generations. In particular, the frequencies of Gofa (GF), Gamo (GM), Gurage (GR) and Wolayta (WL), are substantially lower in the donor's generation (0.7%, 1.8%, 1.3% and 1.2% respectively), than for the father's (1.9%, 2.5%, 2.2%, 2.0%), mother's (2.0%, 2.9%, 2.3% and 1.8% respectively), paternal grandfather's (1.9%, 2.6%, 2.4% and 2.1% respectively), and maternal grandmother's (2.1%, 3.2%, 2.6% and 1.8% respectively) generation.

Table 4.4 Diversity values for first language of the sample donor, donor's parents, paternal grandfather and maternal grandmother in ethnic groups of the sample donor

Ethnic	Donor	s.d	Father	s.d	Mother	s.d	Paternal	s.d	Maternal	s.d
AA	0.225	0.039	0.161	0.034	0.147	0.033	0.116	0.030	0.116	0.030
AF	0.018	0.013	0.018	0.013	0.120	0.031	0.035	0.017	0.137	0.032
AG	0.497	0.030	0.439	0.030	0.404	0.030	0.417	0.030	0.417	0.030
AL	0.137	0.033	0.071	0.024	0.354	0.046	0.089	0.027	0.383	0.046
AM	0.010	0.005	0.134	0.017	0.117	0.016	0.175	0.019	0.150	0.018
AN	0.000	0.000	0.019	0.013	0.000	0.000	0.037	0.018	0.000	0.000
BE	0.032	0.016	0.016	0.011	0.094	0.026	0.032	0.016	0.109	0.028
BK	0.069	0.024	0.120	0.031	0.054	0.021	0.136	0.032	0.072	0.024
BN	0.031	0.015	0.000	0.000	0.150	0.032	0.016	0.011	0.150	0.032
BR	0.409	0.045	0.050	0.020	0.145	0.032	0.017	0.012	0.160	0.034
BU	0.275	0.040	0.248	0.038	0.286	0.040	0.248	0.038	0.310	0.041
DI	0.215	0.036	0.250	0.038	0.102	0.026	0.199	0.035	0.074	0.023
DR	0.510	0.048	0.192	0.038	0.347	0.046	0.210	0.039	0.420	0.047
DS	0.056	0.022	0.056	0.022	0.163	0.036	0.038	0.019	0.180	0.037
DW	0.130	0.031	0.051	0.020	0.132	0.031	0.067	0.023	0.194	0.037
DZ	0.094	0.029	0.076	0.026	0.076	0.026	0.076	0.026	0.076	0.026
GB	0.332	0.044	0.319	0.044	0.365	0.045	0.327	0.044	0.356	0.045
GE	0.016	0.011	0.033	0.016	0.016	0.011	0.033	0.016	0.000	0.000
GF	0.560	0.047	0.310	0.044	0.484	0.047	0.326	0.044	0.427	0.047
GG	0.000	0.000	0.158	0.035	0.175	0.036	0.158	0.035	0.175	0.036
GM	0.451	0.034	0.625	0.033	0.637	0.033	0.618	0.034	0.631	0.033
GN	0.453	0.047	0.104	0.029	0.201	0.038	0.137	0.032	0.201	0.038
GR	0.513	0.041	0.426	0.040	0.453	0.040	0.360	0.039	0.379	0.039
GW	0.017	0.012	0.051	0.020	0.034	0.017	0.017	0.012	0.034	0.017
HD	0.281	0.040	0.177	0.034	0.357	0.043	0.191	0.035	0.432	0.044
HM	0.086	0.026	0.053	0.021	0.121	0.031	0.035	0.017	0.104	0.029
KF	0.212	0.037	0.216	0.038	0.097	0.027	0.204	0.037	0.081	0.025
KM	0.203	0.037	0.067	0.023	0.328	0.043	0.132	0.031	0.353	0.044
KN	0.000	0.000	0.126	0.032	0.073	0.025	0.143	0.034	0.091	0.028
KR	0.498	0.048	0.019	0.013	0.333	0.045	0.000	0.000	0.333	0.045
KS	0.375	0.044	0.190	0.036	0.262	0.040	0.205	0.037	0.247	0.039
MG	0.000	0.000	0.017	0.012	0.034	0.017	0.035	0.017	0.034	0.017
MH	0.207	0.036	0.211	0.036	0.265	0.039	0.227	0.037	0.290	0.040
ML	0.129	0.031	0.159	0.034	0.287	0.041	0.160	0.034	0.300	0.042
NR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
OR	0.402	0.040	0.173	0.031	0.283	0.037	0.184	0.032	0.283	0.037
SC	0.108	0.028	0.094	0.026	0.208	0.036	0.078	0.024	0.221	0.037
SD	0.223	0.037	0.093	0.026	0.164	0.033	0.108	0.028	0.166	0.033
SK	0.018	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SM	0.157	0.035	0.175	0.037	0.193	0.038	0.124	0.032	0.209	0.039
TG	0.523	0.061	0.384	0.060	0.539	0.061	0.360	0.059	0.509	0.062
TY	0.018	0.012	0.035	0.017	0.069	0.024	0.035	0.017	0.069	0.024
WL	0.554	0.047	0.286	0.043	0.446	0.047	0.249	0.041	0.458	0.048
YM	0.019	0.013	0.055	0.022	0.019	0.013	0.037	0.018	0.037	0.018
ZS	0.196	0.038	0.088	0.027	0.172	0.036	0.105	0.029	0.187	0.037

From Table 4.4, the most striking feature is that the diversity values for first language can be quite different between generations for a given ethnic group. This is apparent with instances of the sample donor's diversity of first language being substantially higher than that found in the donor's father, mother, paternal grandfather or maternal grandmother generations, as was observed in the Burji (BR, 0.409, compared to 0.05, 0.145, 0.017 and 0.160 respectively), Kore (KR, 0.498, compared to 0.019, 0.333, 0.000 and 0.333 respectively), Genta (GN, 0.453, compared to 0.104, 0.201, 0.137 and 0.201 respectively), Dirasha (DR, 0.51 compared to 0.192, 0.347, 0.210 and 0.420 respectively) and Oromo (OR, 0.402 compared to 0.173, 0.283, 0.184 and 0.283 respectively). There were also instances where the diversity of the sample donor is substantially lower than that found for the donor's father, mother, paternal grandfather or maternal grandmother generations, as found in the Konta (KN, 0.00, compared to 0.126, 0.073, 0.143 and 0.091 respectively), Ganjule (GG, 0.00, compared to 0.158, 0.175, 0.158 and 0.175 respectively), Amhara (AM, 0.01, compared to 0.134, 0.117, 0.175 and 0.15 respectively) and the Gamo (GM, 0.451, compared to 0.625, 0.637, 0.618 and 0.631). Interestingly, instances where either parent shows higher diversity than the sample donor, the corresponding grandparent would often also exhibit similar levels of diversity. This is clearly shown in the Afar (AF), Bench (BN), Dasanach (DS), Alaba (AL), Maale (ML), Shekecho (SC), Kembata (KM) and Hadiya (HD) where the diversity of both the donor's mother and maternal grandmother are substantially higher than the diversity of the sample donor, and donor's father and paternal grandfather. Instances where the donor's father and paternal grandfather exhibit greater diversity than the sample donor, donor's mother and maternal grandmother can be clearly seen in the Konta (KN) and Basketo (BK), but this is generally a less common phenomenon than the instances of greater maternal diversity.

4.3 How important is it to collect detailed ethnographic data for ethnic groups used in population genetic studies?

4.3.1 Can language be used as a proxy for ethnicity and vice versa?

Table 4.5 shows the proportion of samples in ethnic groups of the sample donor and the donor's parents, paternal grandfathers and maternal grandmothers who speak the traditional language of their ethnic group as a first language (the language that is eponymous with, and traditionally spoken by, members of an ethnic group). The most striking feature of this table is differences in the proportion of samples that speak the traditional language of the ethnic group in the sample donor's generation. The mean for the proportion of Ethiopian sample donors who speak the traditional language of their ethnic group was only 0.838, compared with 0.959, 0.960, 0.977 and 0.976 for the donor's father, mother, paternal grandfather and maternal grandmother respectively. Furthermore, in four of the ethnic groups the proportion of sample donors that speak their traditional language as a first language is under 50%, namely the Gamo (GM), Gofa (GF), Gurage (GR) and Tigray (TG) (values of 0.330, 0.342, 0.454 and 0.485 respectively). In 19 of the 45 ethnic groups, there was a 10% decrease in the proportion of donors who speak the traditional language of their ethnic group compared to the mean of their parents. The decrease in the numbers of donors speaking the traditional language of their ethnic group in the sample donor's generation coincides with an increase in the number speaking Amhara as a first language. The frequency of Amhara as a first language was 19.8%, 11.0%, 11.9%, 9.9%, and 10.8%, (in the donor's, donor's father's, mother's, paternal grandfather's and maternal grandmother's generation respectively) compared to the frequencies of Amhara as an ethnicity: 6.9%, 8.1%, 9.0%, 8.3% and 9.2% in the donor's, donor's father's, mother's, paternal grandfather's and maternal grandmother's generation respectively (Supplementary Table Language, Supplementary Table Ethnic).

Table 4.5 Proportion of samples in each generation that speak the traditional language of their ethnic group as a first language

Ethnic group	Donor	Father	Mother	Paternal	Maternal
AA	0.872	0.974	0.969	0.991	0.992
AF	0.991	0.991	1.000	1.000	1.000
AG	0.565	0.735	0.824	0.780	0.824
AL	0.927	1.000	0.967	1.000	1.000
AM	0.992	0.991	0.986	0.998	0.989
AN	1.000	1.000	1.000	0.991	1.000
BE	0.984	0.992	0.992	0.992	1.000
BK	0.965	1.000	1.000	1.000	1.000
BN	0.984	1.000	1.000	1.000	1.000
BR	0.723	0.983	1.000	1.000	1.000
BU	0.849	0.940	0.932	0.957	0.956
DI	0.879	1.000	0.992	1.000	1.000
DR	0.648	0.971	0.990	1.000	1.000
DS	0.971	0.990	1.000	1.000	1.000
DW	0.932	1.000	1.000	0.992	0.991
DZ	0.952	0.990	1.000	1.000	1.000
GB	0.805	0.836	0.829	0.860	0.880
GE	0.992	1.000	0.992	1.000	1.000
GF	0.342	0.955	0.974	0.982	0.983
GG	1.000	1.000	1.000	1.000	1.000
GM	0.330	0.789	0.751	0.912	0.887
GN	0.717	1.000	1.000	1.000	1.000
GR	0.454	0.810	0.821	0.904	0.936
GW	0.991	0.983	0.992	1.000	1.000
HD	0.835	0.983	0.992	0.983	0.992
HM	0.955	0.982	0.991	0.991	0.991
KF	0.883	0.991	1.000	1.000	1.000
KM	0.889	1.000	0.991	1.000	1.000
KN	1.000	1.000	1.000	1.000	1.000
KR	0.676	0.991	1.000	1.000	0.989
KS	0.767	0.991	0.982	1.000	1.000
MG	1.000	1.000	1.000	1.000	1.000
MH	0.885	0.913	0.917	0.943	0.925
ML	0.933	0.956	0.990	0.991	0.990
NR	1.000	1.000	1.000	1.000	1.000
OR	0.725	0.953	0.874	0.981	0.956
SC	0.944	1.000	1.000	1.000	1.000
SD	0.873	1.000	1.000	0.992	0.992
SK	0.991	1.000	1.000	1.000	1.000
SM	0.917	0.980	0.980	1.000	0.979
TG	0.485	0.917	0.937	0.974	0.957
TY	0.991	1.000	0.991	1.000	0.991
WL	0.582	0.934	0.927	1.000	0.971
YM	0.991	0.991	1.000	1.000	1.000
ZS	0.892	0.991	1.000	0.991	1.000
Ethiopia mean	0.838	0.959	0.960	0.977	0.976

Figure 4.3 Plot of the proportion of individuals in each generation that speak the traditional language of their ethnic group.

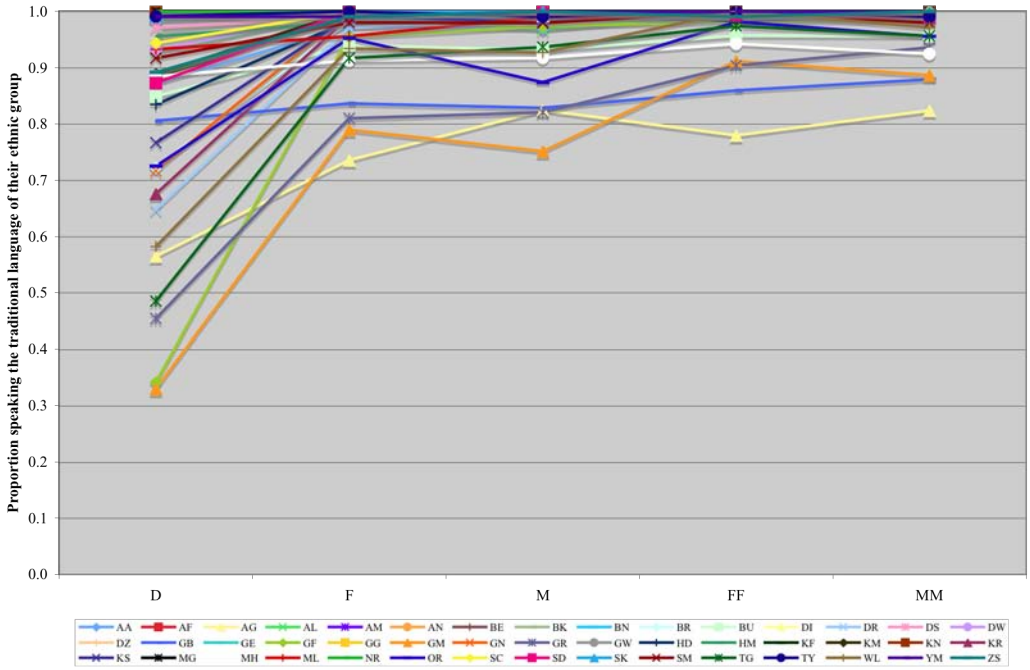
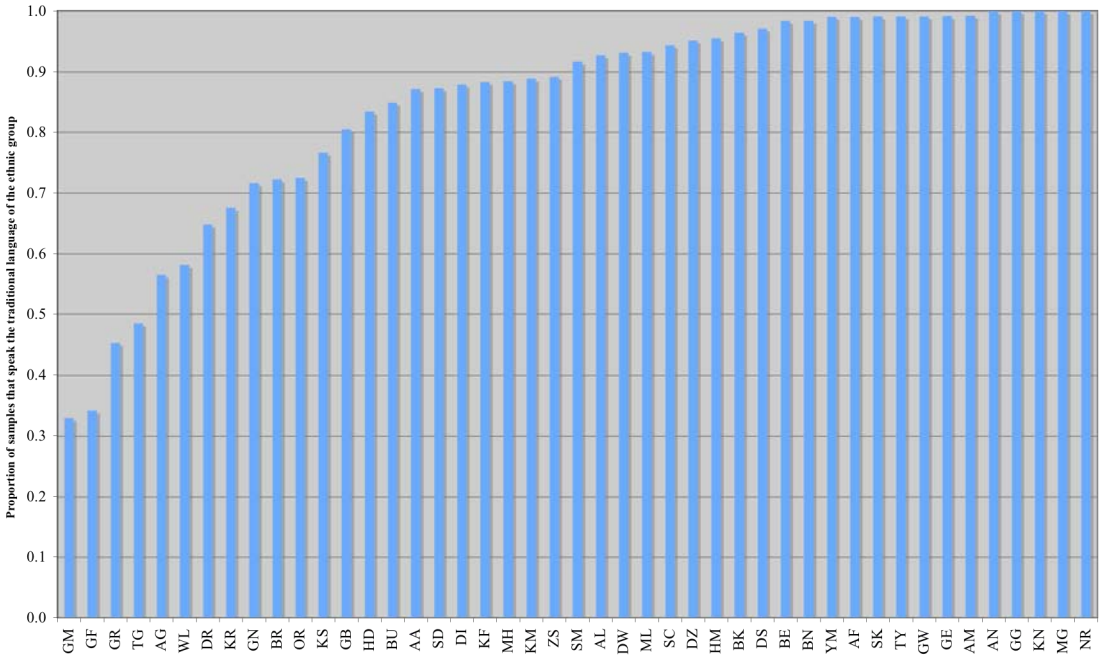


Figure 4.4 Proportion of sample donors that speak the traditional language of their ethnic group (ordered by increasing value)



To see how this degree of discordance between ethnicity and traditional language can affect genetic analysis based upon linguistics, AMOVA was performed using both samples assigned to their traditional language of their ethnic group (Table 4.6) and samples assigned to groups using their actual first language (Table 4.7), and then grouped by linguistic group (Cushitic, Omotic, Nilo-Saharan, Semitic). In all cases, the amount of variance attributable to differences between populations (Global Fst), be they ethnic groups or individual languages, was highly significant ($p<0.01$) using all metrics, with similar Global Fst values for a given metric in the two data sets. However, the amount of variance attributable to differences among linguistic groups (Fct) showed some clear differences when samples were assigned to either their traditional language or actual first language. For AMOVA performed using traditional languages, all metrics for among linguistic group variances were significant, whereas for AMOVA performed using first languages spoken, significant among linguistic group variances were only observed for mtDNA K2P and NRY MS Rst, and were lower values than were observed for these metrics using traditional languages. In all cases the amount of variance attributable to differences within linguistic groups (Fsc) was greater than that attributable to differences between linguistic groups (Fct). This difference in among linguistic group variances using the two datasets is likely due to the lower proportion of sample donor's speaking the traditional language of their ethnic groups (in all linguistic groups) compared to previous generations, with a resultant increase in non-ethnic Amhara samples speaking Amhara as a first language. As a consequence analysis performed using first language groupings demonstrate an overall decrease in the amount of variance attributable to differences between linguistic groups as many traditionally Omotic, Cushitic and Nilo-Saharan speakers are grouped with Semitic speakers in the analysis.

Table 4.6 AMOVA results, performed using samples assigned to ethnic groups, grouped by the Linguistic group of the ethnic group's traditional language

	Fsc	Fct	Global Fst (ungrouped data)
mtDNA Fst	0.008 (<0.001)	0.002 (<0.001)	0.010 (<0.001)
mtDNA K2P	0.016 (<0.001)	0.008 (<0.001)	0.021 (<0.001)
NRY UEP Fst	0.108 (<0.001)	0.037 (0.003)	0.130 (<0.001)
NRY UEP-MS Fst	0.054 (<0.001)	0.007 (0.005)	0.058 (<0.001)
NRY MS Rst	0.092 (<0.001)	0.052 (<0.001)	0.124 (<0.001)

Table 4.7 AMOVA results, performed using samples assigned to first languages, grouped by the Linguistic group of their first language

	Fsc	Fct	Global Fst (ungrouped data)
mtDNA Fst	0.009 (<0.001)	**<0.001 (0.070)**	0.009 (<0.001)
mtDNA K2P	0.017 (<0.001)	0.004 (0.003)	0.020 (<0.001)
NRY UEP Fst	0.117 (<0.001)	**0.018 (0.039)**	0.128 (<0.001)
NRY UEP-MS Fst	0.059 (<0.001)	**-0.003 (0.577)**	0.057 (<0.001)
NRY MS Rst	0.104 (<0.001)	0.031 (0.002)	0.124 (<0.001)

4.3.2 Are geographic distances and sex specific genetic distances correlated with linguistic, and ethnic variation in Ethiopia?

Table 4.8 shows the results of Mantel tests of correlation between genetic distances using NRY and mtDNA haplotypes, linguistic distances based on the frequencies of first languages (Supplementary Table LangDist), and ethnic distances based on the frequency of different ethnicities in the parents and grandparents of sample donors (Supplementary Table EthDist). Geographic distance (using the weighted mean collection location for an ethnic group) was found to correlate highly significantly ($p=0.0002$, $r=0.5067$) with mtDNA genetic distances when the similarity between haplotypes is taken into account (as determined by K2P), but was not significantly correlated ($p>0.05$) when distances were estimated purely on the frequency of the mtDNA haplotypes (Fst). Similarly, geographic distance was significantly correlated ($p=0.0205$, $r=0.2778$) with NRY genetic distances when the similarity between

haplotypes is taken into account (as estimated by Rst distance using microsatellite allele sizes), but was not significantly correlated when distances were based on haplotype frequencies alone (Fst, $p>0.05$). Furthermore, when the deeper evolutionary similarity between the NRY haplotypes in ethnic groups was taken into account using the frequencies of the UEP haplogroups in ethnic groups (Fst), a stronger correlation was observed between NRY genetic distance and geographic distance ($p=0.0128$, $r=0.2885$), than was observed for NRY Rst distances. Interestingly, no significant correlations with geographic distance were observed between any of the linguistic distances or ethnic distances for non donor generations as determined by the frequencies of first language and ethnicity respectively (Fst in both cases).

Linguistic distance based on the first language of the sample donor was shown to be significantly correlated ($p<0.01$) with all estimates of NRY and mtDNA genetic distances, as well as all non donor generation linguistic and ethnic distances. There was a strong correlation between all pairs of non donor generation linguistic and ethnic distances ($p<0.0001$, $r>0.6000$). Maternal (mother and maternal grandmother) linguistic and ethnic distances were significantly correlated with all genetic distances ($p<0.01$ in all cases except for NRY UEP-MS Fst where $p<0.05$). Paternal (father and paternal grandfather) linguistic distances were significantly correlated with all NRY genetic distances, with a more significant correlation with NRY MS Rst ($p<0.01$), than with NRY UEP Fst and NRY UEP-MS Fst (where $p<0.05$). mtDNA genetic distances did not generally correlate with paternal linguistic distances except in one instance where the linguistic distances using the frequency of the paternal grandfather's first language was weakly correlated with mtDNA Fst ($p=0.0320$). Paternal ethnic distance was significantly correlated with NRY MS Rst ($p<0.01$). A weaker correlation was observed between the ethnic distance based on ethnicity of the donor's father and NRY UEP Fst ($p=0.0257$) than for ethnic distance based on the ethnicity of the paternal grandfather ($p=0.0053$). Similarly, NRY UEP-MS Fst did not correlate with the ethnic distance based on the ethnicity of the donor's father ($p>0.05$), but a weak correlation was observed with the ethnic distance based on the ethnicity of the donor's paternal grandfather ($p=0.0165$). Interestingly, a significant correlation was observed between all paternal ethnic distances and all mtDNA genetic distances ($p<0.05$, correlation between the ethnic distance using the paternal grandfather's ethnicity and mtDNA Fst, $p<0.01$).

Table 4.8 Summary of Mantel test results using genetic, linguistic and ethnic distance between ethnic groups in different generations of the sample donor (Mantel r values in lower diagonal, p values in upper diagonal ($p<0.01$, $p<0.05$))

Distances	Geographic	NRY UEP Fst	NRY UEP-MS Fst	NRY MS Rst	mtDNA HVS1 Fst	mtDNA HVS1 K2P	Linguistic: Donor	Linguistic: Father	Linguistic: Mother	Linguistic: Paternal grandfather	Linguistic: Maternal grandmother	Ethnic: Father	Ethnic: Mother	Ethnic: Paternal grandfather	Ethnic: Maternal grandmother
Geographic	*	0.0128	0.2670	0.0205	0.0614	0.0002	0.1626	0.5493	0.2309	0.4860	0.2582	0.1525	0.1451	0.1470	0.1209
NRY UEP Fst	0.2885	*	0.1435	<0.0001	0.0072	0.0012	0.0010	0.0187	0.0008	0.0129	0.0005	0.0257	0.0006	0.0053	0.0003
NRY UEP-MS Fst	0.0644	0.1304	*	0.0015	0.0037	0.0053	0.0074	0.0110	0.0099	0.0112	0.0171	0.0565	0.0320	0.0165	0.0382
NRY MS Rst	0.2778	0.6915	0.4288	*	0.0005	<0.0001	0.0012	0.0011	<0.0001	0.0014	<0.0001	0.0072	<0.0001	0.0016	<0.0001
mtDNA HVS1 Fst	0.2054	0.3065	0.3968	0.3919	*	<0.0001	0.0004	0.0546	0.0003	0.0320	0.0002	0.0270	<0.0001	0.0071	<0.0001
mtDNA HVS1 K2P	0.5067	0.3914	0.3994	0.4630	0.6551	*	0.0031	0.1076	0.0066	0.0823	0.0055	0.0249	0.0017	0.0124	0.0011
Linguistic: Donor	0.1320	0.2772	0.3080	0.3279	0.3988	0.2862	*	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Linguistic: Father	0.0007	0.1961	0.2646	0.2917	0.2026	0.1463	0.7142	*	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Linguistic: Mother	0.0971	0.2935	0.2971	0.4029	0.3927	0.2631	0.8326	0.8118	*	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Linguistic: Pat.Gfather	0.0190	0.2100	0.2701	0.2996	0.2328	0.1563	0.6948	0.9837	0.8243	*	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Linguistic: Mat.Gmother	0.0907	0.3124	0.2760	0.4143	0.4118	0.2683	0.7940	0.7657	0.9848	0.7920	*	<0.0001	<0.0001	<0.0001	<0.0001
Ethnic: Father	0.1219	0.1851	0.1968	0.2573	0.2349	0.2031	0.5616	0.8253	0.6736	0.8294	0.6361	*	<0.0001	<0.0001	<0.0001
Ethnic: Mother	0.1403	0.3078	0.2423	0.4001	0.4447	0.2949	0.7474	0.6237	0.9121	0.6512	0.9156	0.7004	*	<0.0001	<0.0001
Ethnic: Pat.Gfather	0.1263	0.2294	0.2470	0.2959	0.2889	0.2238	0.6106	0.8599	0.7537	0.8868	0.7268	0.9745	0.7537	*	<0.0001
Ethnic: Mat.Gmother	0.1526	0.3256	0.2312	0.4201	0.4636	0.3082	0.7320	0.6357	0.9276	0.6714	0.9520	0.6710	0.9810	0.7406	*

4.4 What are the implications if sample sets are comprised of donors with diverse ethnic ancestries?

The data were filtered to remove donors with parents, paternal grandfathers and maternal grandmothers whose ethnicity did not match that of the sample donor. In total, 950 (16.5%) samples had a non-uniform ethnic ancestry and were removed from the analysis, with the remaining 4806 samples (hereafter referred to as the DFMFFMM sample set) representing 83.5% of the dataset. Two ethnic groups were completely homogeneous in their ethnic ancestry (Nuer (NR) and Sheko (SK)), whereas the Gamo (GM) had 62.2% of samples removed due to one or more donors having parents, paternal grandfathers or maternal grandmothers with ethnicity that did not match that of the donor (Table 4.10).

Table 4.9 Summary of the genetic diversity found within Ethiopian ethnic groups after removal of samples with non-uniform ethnic ancestry.

Ethnic group	n	NRY haplogroup *h*	s.d ±	NRY haplotype *h*	s.d ±	NRY MS MSV	mtDNA haplotype *h*	s.d ±	mtDNA haplotype π	s.d ±
AA	105	0.163	0.036	0.818	0.038	0.472	0.986	0.012	0.0243	0.0125
AF	102	0.574	0.049	0.941	0.023	0.858	0.987	0.011	0.0246	0.0126
AG	218	0.612	0.033	0.954	0.014	1.215	0.989	0.007	0.0233	0.0119
AL	83	0.660	0.052	0.959	0.022	1.126	0.992	0.010	0.0260	0.0133
AM	337	0.611	0.027	0.973	0.009	1.240	0.991	0.005	0.0232	0.0119
AN	107	0.738	0.042	0.911	0.028	0.856	0.992	0.008	0.0267	0.0119
BE	117	0.300	0.042	0.904	0.027	0.470	0.984	0.012	0.0234	0.0120
BK	99	0.265	0.044	0.788	0.041	0.583	0.987	0.012	0.0218	0.0113
BN	117	0.358	0.044	0.919	0.025	0.684	0.984	0.012	0.0219	0.0113
BR	108	0.333	0.045	0.931	0.024	0.867	0.986	0.011	0.0243	0.0124
BU	103	0.293	0.045	0.910	0.028	0.446	0.979	0.014	0.0234	0.0120
DI	108	0.300	0.044	0.908	0.028	0.507	0.975	0.015	0.0227	0.0117
DR	74	0.419	0.057	0.956	0.024	0.681	0.989	0.012	0.0254	0.0131
DS	95	0.251	0.044	0.933	0.026	0.726	0.991	0.010	0.0273	0.0139
DW	102	0.406	0.049	0.951	0.021	0.683	0.993	0.008	0.0238	0.0122
DZ	97	0.239	0.043	0.910	0.029	0.589	0.984	0.013	0.0263	0.0134
GB	99	0.190	0.039	0.937	0.024	0.596	0.974	0.016	0.0225	0.0116
GE	120	0.568	0.045	0.919	0.025	0.941	0.984	0.011	0.0246	0.0126
GF	62	0.436	0.063	0.968	0.022	0.932	0.995	0.009	0.0234	0.0121
GG	97	0.100	0.030	0.785	0.042	0.364	0.967	0.018	0.0255	0.0130
GM	79	0.333	0.053	0.940	0.027	0.798	0.995	0.008	0.0245	0.0126
GN	94	0.199	0.041	0.904	0.030	0.564	0.989	0.011	0.0268	0.0137
GR	113	0.626	0.046	0.969	0.016	1.174	0.990	0.010	0.0231	0.0119
GW	115	0.395	0.046	0.957	0.019	0.449	0.971	0.016	0.0241	0.0124
HD	86	0.408	0.053	0.705	0.049	0.601	0.993	0.009	0.0252	0.0129
HM	105	0.308	0.045	0.929	0.025	0.485	0.979	0.014	0.0238	0.0122
KF	103	0.658	0.047	0.963	0.019	1.307	0.967	0.017	0.0212	0.0110
KM	83	0.562	0.054	0.948	0.024	1.012	0.995	0.008	0.0250	0.0128
KN	96	0.474	0.051	0.949	0.022	0.625	0.993	0.008	0.0228	0.0118
KR	88	0.269	0.047	0.833	0.040	0.479	0.991	0.010	0.0248	0.0127
KS	98	0.291	0.046	0.930	0.026	0.369	0.980	0.014	0.0260	0.0133
MG	111	0.215	0.039	0.728	0.042	0.341	0.963	0.018	0.0232	0.0119
MH	115	0.295	0.043	0.948	0.021	0.759	0.976	0.014	0.0240	0.0123
ML	97	0.243	0.044	0.938	0.025	0.665	0.985	0.012	0.0243	0.0125
NR	118	0.691	0.043	0.940	0.022	0.838	0.990	0.009	0.0264	0.0135
OR	114	0.614	0.046	0.965	0.017	1.273	0.994	0.007	0.0239	0.0123
SC	106	0.604	0.048	0.949	0.021	1.301	0.981	0.013	0.0214	0.0111
SD	113	0.355	0.045	0.948	0.021	0.890	0.993	0.008	0.0254	0.0130
SK	113	0.136	0.032	0.882	0.030	0.382	0.982	0.012	0.0246	0.0126
SM	90	0.511	0.053	0.877	0.035	0.996	0.981	0.014	0.0230	0.0119
TG	35	0.671	0.079	0.963	0.032	1.343	0.990	0.017	0.0237	0.0124
TY	110	0.287	0.043	0.884	0.030	0.503	0.983	0.012	0.0236	0.0121
WL	73	0.499	0.059	0.967	0.021	1.007	0.996	0.008	0.0246	0.0127
YM	104	0.561	0.049	0.905	0.029	0.988	0.978	0.014	0.0247	0.0127
ZS	97	0.211	0.041	0.907	0.029	0.501	0.958	0.020	0.0247	0.0127

Table 4.9 summarises the diversity values in ethnic groups estimated after the removal of samples with non-uniform across generation ethnic identity while Table 4.10 details the resultant changes in these values compared to values estimated from the complete dataset. For gene diversity (*h*) estimated from the frequencies of NRY UEP haplogroups, the mean absolute change in diversity was 0.0166 and the median absolute change was 0.0096. There were 23 ethnic groups registering a decrease in diversity, while the diversity of 20 ethnic groups increased after the removal of samples, which was not a significant deviation from an equal likelihood of either positive or negative difference from the original data (2 tailed Sign test p=0.761). The largest increase in NRY UEP gene diversity was found in the Gofa (GF), which increased by 0.022 after removal of 44.1% of samples, whilst the largest decrease in diversity occurred in the Gamo (GM) with a decrease of 0.105 after removal of 62.2% of samples. For gene diversity estimated from frequencies of NRY UEP-MS haplotypes, the mean absolute change in diversity was 0.0074 and the median absolute change was 0.0051. 33 ethnic groups registered a decrease in diversity and 10 ethnic groups registered an increase in diversity, which was a highly significant deviation from an expectation of equally likely positive or negative differences from the original data (2 tailed Sign test p<0.001). The largest increase in diversity was observed in the Oromo (OR), which increased, by 0.006 after removal of 23.5% of samples, while the largest decrease in diversity was 0.038 after removal of 32.3% of samples in the Hadiya (HD). For NRY MSV, the mean absolute change was 0.0259 and the median change was 0.0136. There were 28 ethnic groups decreasing in MSV, whilst the MSV in 15 ethnic groups increased in comparison with the original data, although this was not significantly different from chance (2 tailed Sign test p=0.066). The largest increase in MSV was 0.028 and occurred in the Kefa (KF) after removal of 14.2% of samples, whilst the largest decrease in MSV was 0.119 and was observed in the Kore (KR) after removal of 18.5% of samples. For mtDNA HVS1 haplotype gene diversity, the mean absolute change in diversity was 0.0011 and the median change was 0.0008. There were 26 ethnic groups decreasing in diversity, whilst 17 ethnic groups registering increases in diversity after removal of samples, but this was no different from chance (2 tailed Sign test p=0.222). The largest increase in diversity was 0.0024 and occurred in the Gofa (GF), whilst the largest decrease in diversity was 0.0064, and was observed in the Konso (KS) after removal of 18.3% of samples. For mtDNA nucleotide diversity, the mean absolute change in diversity was 0.00023 and the median value was 0.00020. Nucleotide diversity decreased in 23 ethnic groups and increased in 20 ethnic groups compared to

the original data, which was not a significant deviation from chance (2 tailed Sign test p=0.761). The largest increase in diversity was 0.0007, and occurred in the Genta (GN) after removal of 16.8% of samples and the largest decrease in diversity was observed in the Kembata (KM), with a decrease of 0.0008 after removal of 29.1% of samples.

Table 4.10 Summary of the change in diversity metrics within Ethiopian ethnic groups after removal of samples with non-uniform ethnic identity across generations compared with the values estimated using the complete dataset.

Ethnic group	n	Proportion of samples with uniform ethnic ancestry	Difference in NRY UEP *h*	Difference in NRY UEP-MS *h*	Difference in NRY MS MSV	Difference in mtDNA HVS1 *h*	Difference in mtDNA HVS1 π
AA	105	0.897	-0.058	-0.021	-0.054	-0.0003	0.0002
AF	102	0.911	0.014	0.001	0.007	0.0004	0.0000
AG	218	0.810	0.001	0.000	0.004	0.0003	-0.0002
AL	83	0.755	0.016	-0.009	0.004	-0.0011	0.0003
AM	337	0.851	-0.005	0.001	-0.010	0.0010	0.0000
AN	107	0.991	0.002	-0.002	0.005	0.0002	0.0001
BE	117	0.944	-0.012	-0.003	-0.014	-0.0006	-0.0003
BK	99	0.876	0.001	0.005	-0.003	-0.0004	-0.0004
BN	117	0.921	-0.054	-0.003	-0.084	-0.0006	-0.0002
BR	108	0.908	0.014	-0.002	0.006	-0.0004	-0.0006
BU	103	0.817	0.021	-0.009	-0.020	-0.0004	0.0005
DI	108	0.818	-0.068	-0.022	-0.093	0.0009	0.0001
DR	74	0.685	-0.010	-0.011	-0.034	-0.0028	0.0003
DS	95	0.905	-0.025	-0.006	-0.018	-0.0012	0.0003
DW	102	0.872	-0.003	0.000	-0.003	-0.0004	0.0000
DZ	97	0.933	0.000	-0.006	-0.014	-0.0009	0.0000
GB	99	0.876	-0.009	-0.001	-0.010	-0.0015	0.0001
GE	120	0.984	-0.002	-0.001	0.004	-0.0005	-0.0001
GF	62	0.559	0.022	-0.001	-0.008	0.0024	-0.0003
GG	97	0.890	-0.023	-0.026	-0.074	-0.0012	-0.0003
GM	79	0.378	-0.105	-0.015	-0.107	0.0003	0.0002
GN	94	0.832	-0.043	-0.014	-0.080	-0.0010	0.0007
GR	113	0.743	-0.012	0.005	0.007	0.0004	0.0003
GW	115	0.983	0.006	0.002	0.005	0.0011	0.0001
HD	86	0.677	0.005	-0.038	0.014	0.0019	0.0000
HM	105	0.938	0.001	-0.003	-0.017	-0.0010	-0.0001
KF	103	0.858	0.010	-0.004	0.028	-0.0050	-0.0005
KM	83	0.709	-0.009	-0.008	-0.019	0.0002	-0.0008
KN	96	0.897	0.014	-0.005	-0.002	0.0004	-0.0005
KR	88	0.815	-0.058	-0.019	-0.119	0.0003	-0.0002
KS	98	0.817	-0.021	-0.016	-0.081	-0.0064	-0.0001
MG	111	0.965	-0.007	-0.017	-0.003	0.0008	-0.0001
MH	115	0.885	0.006	0.002	0.015	0.0002	0.0004
ML	97	0.815	-0.012	0.001	0.002	0.0007	0.0004
NR	118	1.000	0.000	0.000	0.000	0.0000	0.0000
OR	114	0.765	0.002	0.006	-0.035	-0.0008	-0.0002
SC	106	0.848	-0.015	0.002	-0.011	0.0009	-0.0001
SD	113	0.897	-0.016	-0.008	-0.033	-0.0004	0.0000
SK	113	1.000	0.000	0.000	0.000	0.0000	0.0000
SM	90	0.833	-0.003	-0.014	-0.005	-0.0021	0.0001
TG	35	0.530	0.022	-0.006	0.028	-0.0012	0.0003
TY	110	0.965	0.009	-0.002	0.001	-0.0009	-0.0003
WL	73	0.664	0.000	-0.005	-0.060	-0.0005	-0.0002
YM	104	0.963	0.000	-0.006	0.001	-0.0011	-0.0001
ZS	97	0.874	0.009	-0.004	-0.023	-0.0042	0.0005

Table 4.11 Results for rank correlations between absolute difference in diversity metrics and the proportion of samples removed from ethnic groups

Metric	Spearman's r	P value
NRY UEP *h*	0.3629	0.0143
NRY UEP-MS *h*	0.3614	0.0147
NRY MS MSV	0.4583	0.0015
mtDNA HVS1 *h*	0.2202	0.1461
mtDNA HVS1 π	0.3412	0.0218

The results of ranked correlations between the proportion of samples removed from ethnic groups and the resultant change in diversity metrics are shown in Table 4.11, with the data plotted in Figure 4.5 to Figure 4.9. A highly significant correlation ($p=0.0015$, Spearman's $r = 0.4583$) was observed between the proportion of samples removed from ethnic groups and the absolute change in NRY MS MSV values. Significant correlations ($p<0.05$) were also observed between the proportion of samples removed and the absolute changes in both NRY diversity metrics assessed at the haplogroup and haplotype level. Correlations between the proportion of samples removed and the absolute change in mtDNA diversity was not significant when diversity was estimated by the frequency of HVS1 haplotypes, but was significant ($p=0.0218$) when estimated by nucleotide diversity (π).

Figure 4.5 to Figure 4.9 show plots of the absolute differences in the various diversity metrics plotted against the proportion of samples removed in ethnic groups. If the samples that were removed from ethnic groups due to their non-homogeneous ethnicity across generations had a similar distribution of haplotypes to the samples that are retained, then it would be expected that that there would be little change in the diversity metrics due to the removal of samples. If however, the samples that were removed had a significantly different distribution of haplotypes compared to the samples that were retained, then it would be expected that there would be a substantial change in the diversity metrics.

In Figure 4.5 it can be seen that removal of samples does not effect the NRY UEP gene diversity of most of the ethnic groups to a great degree, with the majority of groups clustering near the bottom left of the graph, with a median absolute change in diversity of 0.0096, and ethnic groups such as the Wolayta (WL) exhibiting a change of 0.0005

despite the removal of 33.6% of samples. Five ethnic groups appear to be particularly sensitive to the removal of ancestrally ethnically diverse samples, namely the Bench (BN), Ari (AA), Genta (GN), Kore (KR), and Dizi (DI) groups, appearing towards the upper left of the graph, all with absolute changes in diversity greater than 0.04 with the removal of less than 20.0% of samples in each.

Figure 4.5 Plot of the absolute difference in NRY UEP gene diversity against the proportion of samples removed in ethnic groups

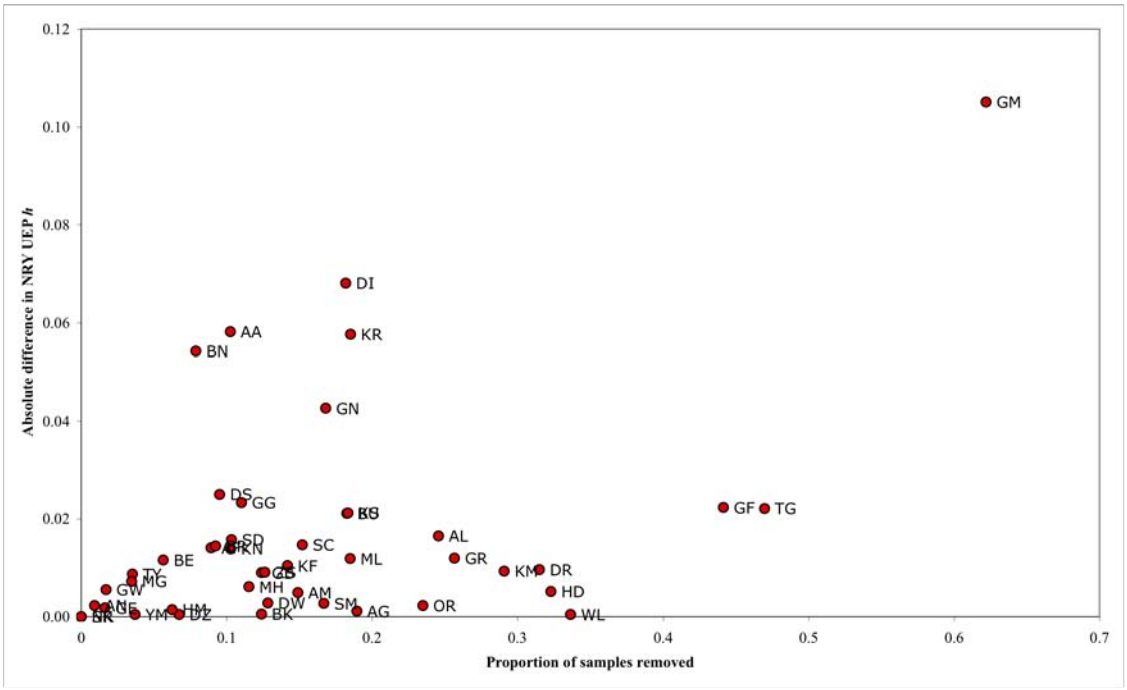
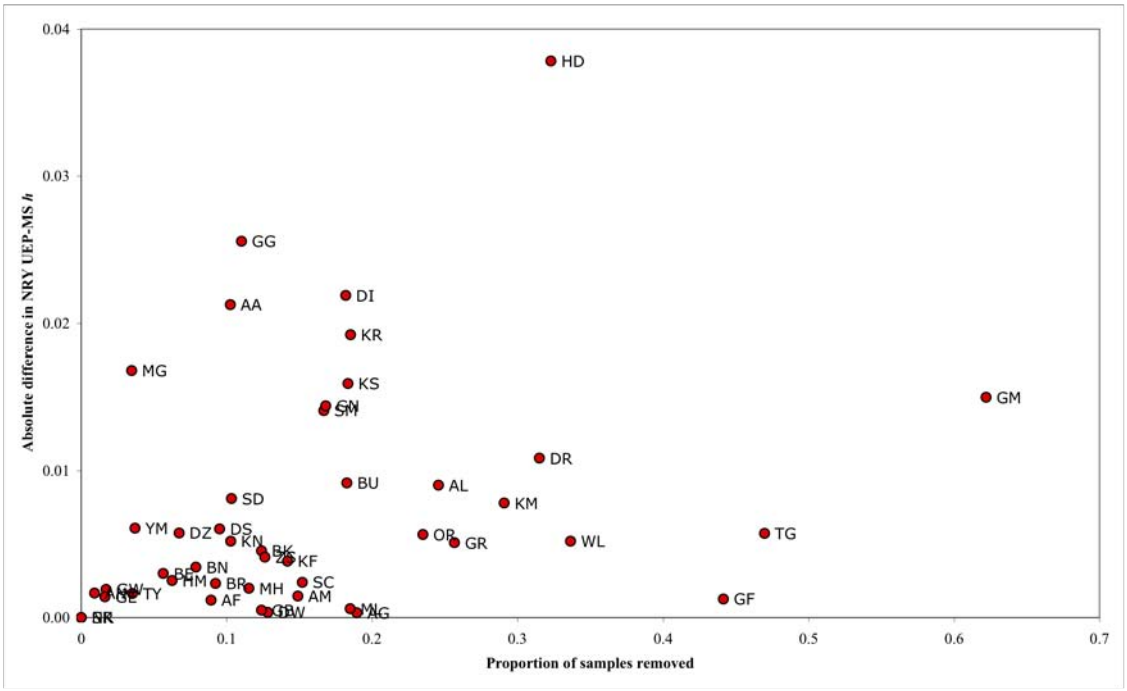


Figure 4.6 shows the majority of ethnic groups in the plot clustering towards the bottom left of the graph, with a median absolute change in NRY UEP-MS gene diversity of 0.0051. Similar to the distribution observed in Figure 4.5, the diversity of ethnic groups such as the Gofa (GF) and the Tigray (TG) seem to be particularly insensitive to the removal of across generational ethnically non-uniform samples. By far the largest absolute change in diversity occurred in the Hadiya (HD), with a difference of 0.0378 after the removal of 32.3% of samples. Ethnic groups that appear to be particular sensitive to the removal of samples with a diverse ethnic across generation difference include the Majenger (MG), Ari (AA) and Ganjule (GG), with the Majenger in particular exhibiting an absolute change in diversity of 0.0168 after the removal of just 3.5% of samples.

Figure 4.6 Plot of the absolute difference in NRY UEP-MS gene diversity against the proportion of samples removed in ethnic groups



In Figure 4.7, the median change in NRY MS MSV after removal of across generation ethnically diverse samples was 0.0136. Nine ethnic groups exhibited an absolute change in MSV of greater than 0.04, with the Bench (BN) appearing to be particularly sensitive to the removal of samples (a decrease in MSV of 0.0841 after removal of 7.9% of samples), indicating the MS haplotypes of the removed samples to be substantially different from the remaining samples. .

Figure 4.7 Plot of the absolute difference in NRY MS MSV against the proportion of samples removed in ethnic groups

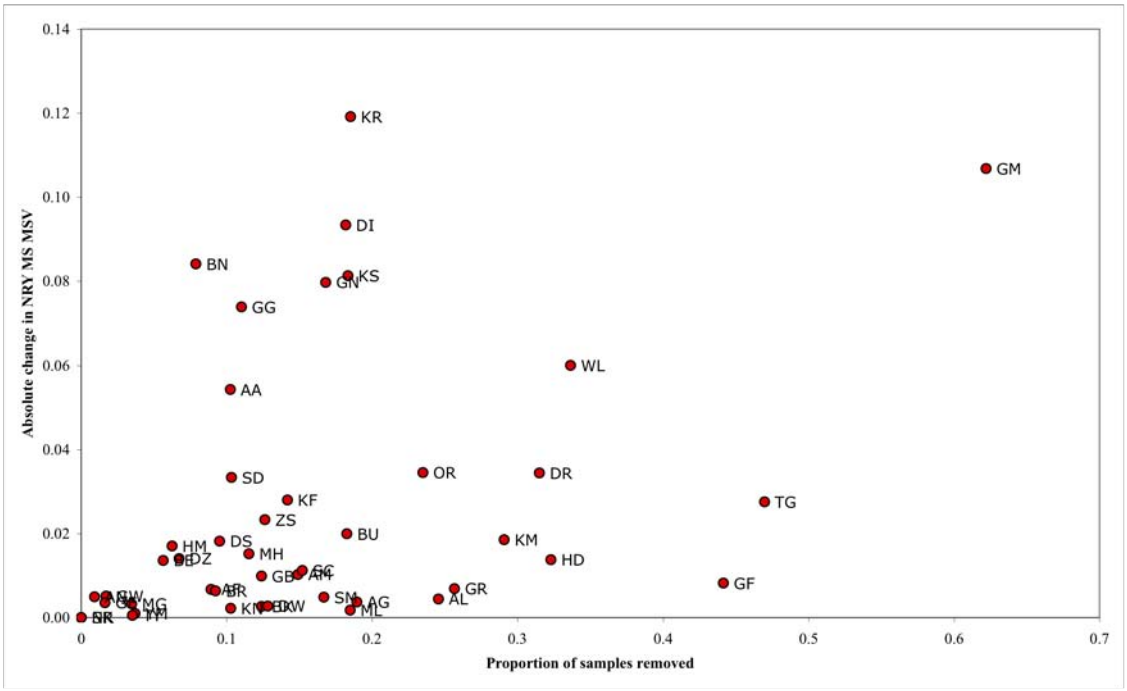
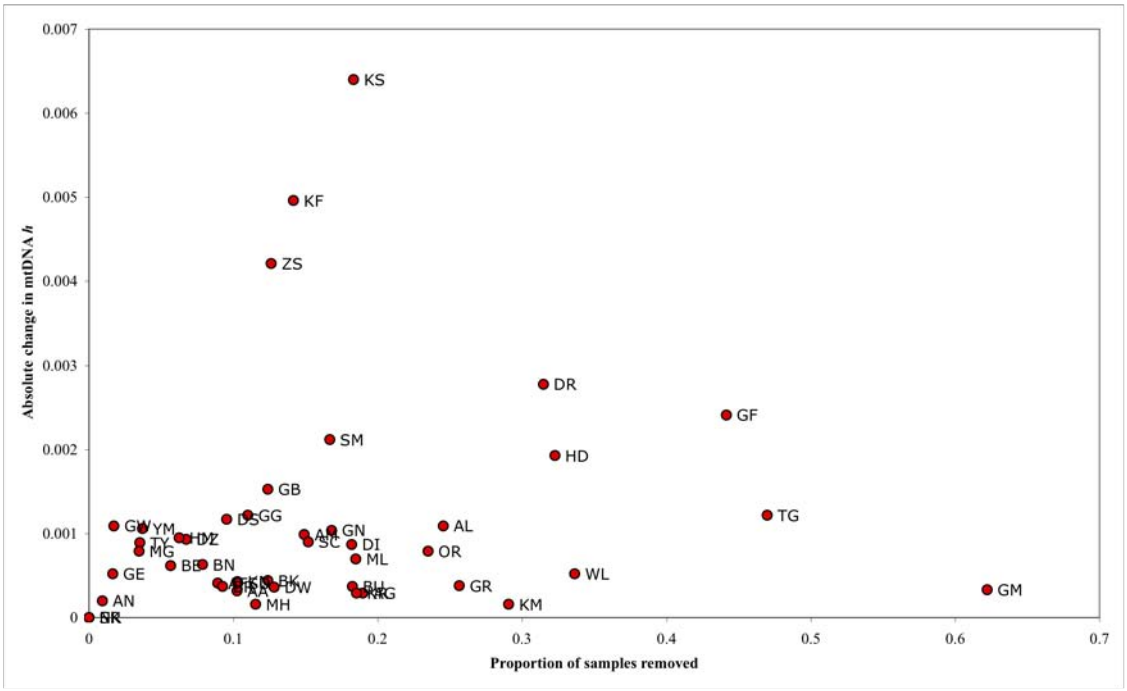


Figure 4.8 shows that generally the mtDNA HVS1 gene diversity of ethnic groups was not greatly affected by the removal of samples with a diverse across generation ethnic identity, with a median absolute change in gene diversity of 0.0008. Exceptions to this however include the Zayse (ZS), Kefa (KF) and in particular the Konso (KS), that exhibited absolute changes in diversity of 0.0042, 0.0046 and 0.0064 respectively after the removal of 12.6%, 14.2% and 18.3% of samples respectively.

Figure 4.8 Plot of the absolute difference in mtDNA HVS1 gene diversity against the proportion of samples removed in ethnic groups



The significant relationship between the removal of samples with a diverse across generation ethnic identity and the change in mtDNA HVS1 nucleotide diversity values can be seen in Figure 4.9. The median change in nucleotide diversity was 0.0002, with eight ethnic groups with absolute changes in diversity of greater than 0.0004. In particular, the Burji (BR), Genta (GN) and Kembata (KM) exhibited absolute changes in diversity of 0.0006, 0.0007 and 0.0008 respectively after the removal of 9.2%, 16.8% and 29.1% respectively. Similar to what was observed for the absolute changes in NRY diversity, the mtDNA diversity of the Gofa (GF) and Tigray (TG) ethnic groups are not particular affected by the removal of substantial numbers of samples with diverse ethnic ancestry (44.1% and 47.0% removed respectively).

Figure 4.9 Plot of the absolute difference in mtDNA HVS1 nucleotide diversity against the proportion of samples removed in ethnic groups

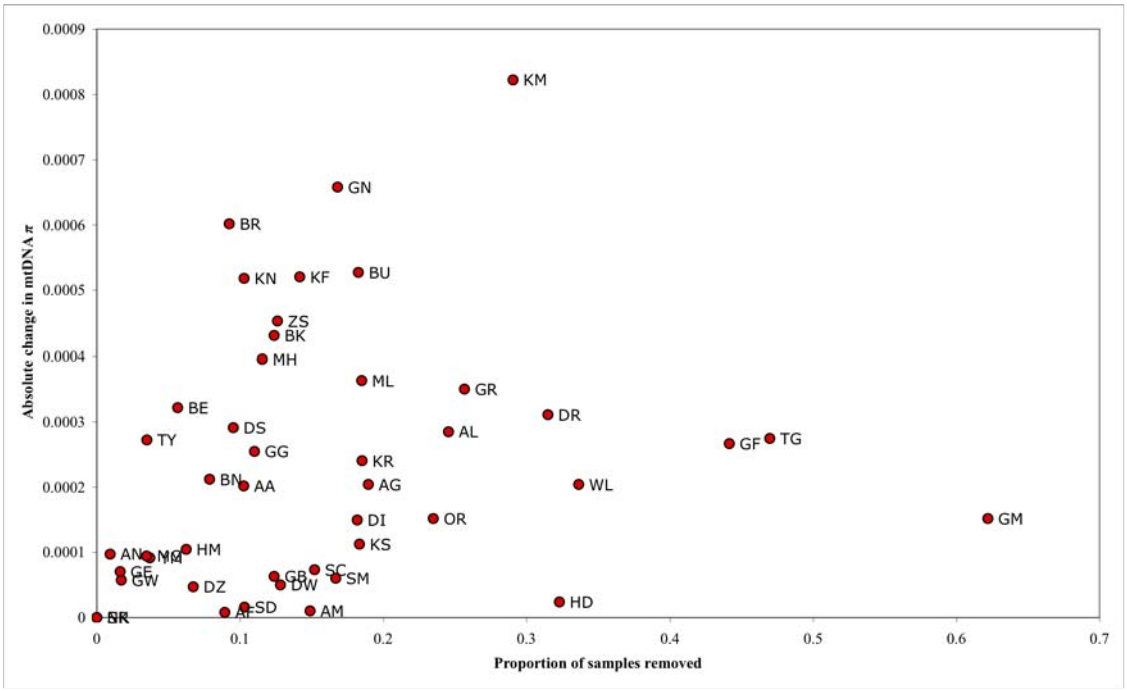
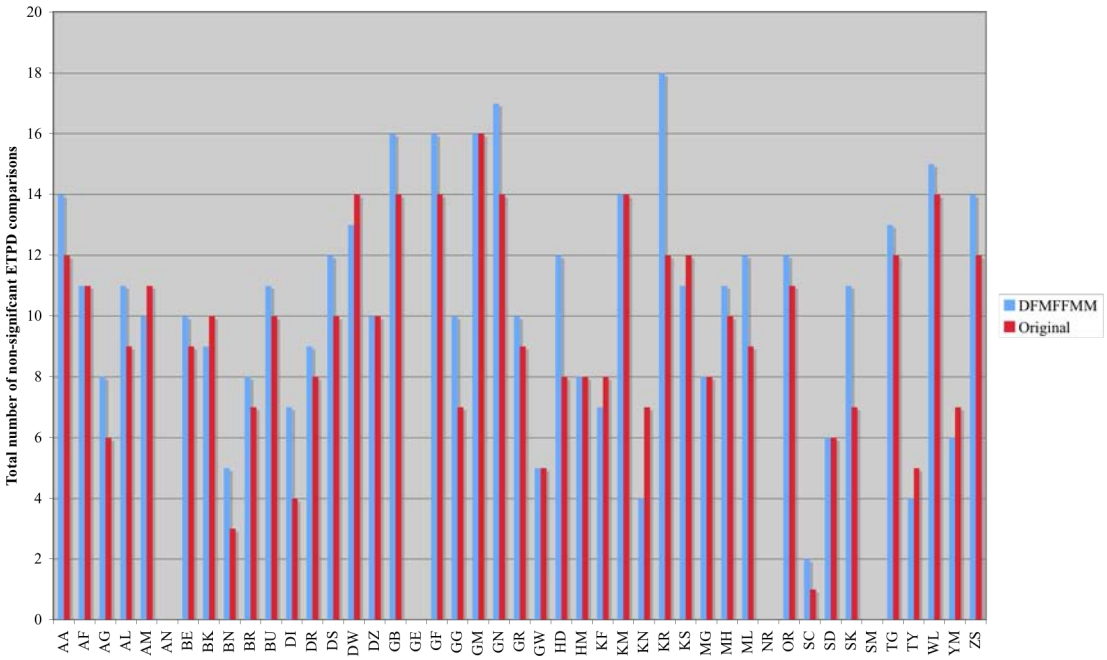


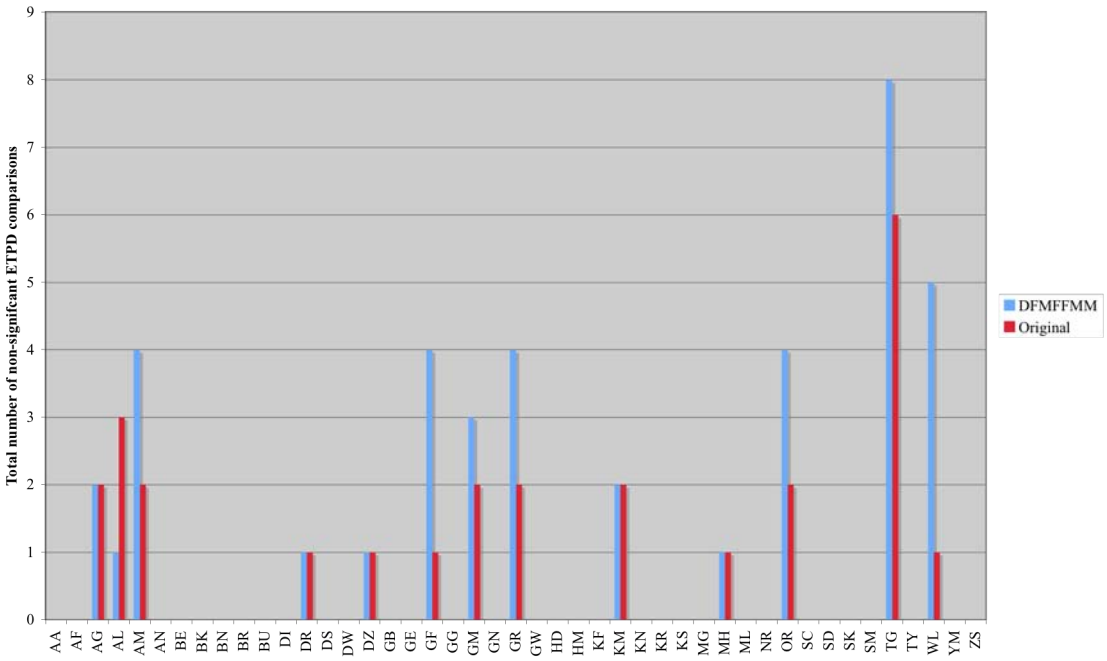
Figure 4.10 to Figure 4.12 show the total number of non-significant ETPD comparisons ($p>0.01$) for the original data and the data that has been filtered for samples with a uniform across generation ethnic identity (Supplementary Table AncEthETPD). It might be expected that the removal of ethnically diverse samples would result in greater distinctiveness in the ethic groups, but this is clearly not the case. In all three figures, removing samples with diverse across generation ethnic identity leads to a increase in the total number of non-significant ETPD comparisons.

Figure 4.10 Total number of non-significant (p>0.01) ETPD comparisons in the original data and across generation ethnicity matched dataset (DFMFFMM), based on the frequency of NRY UEP haplogroups in ethnic groups.



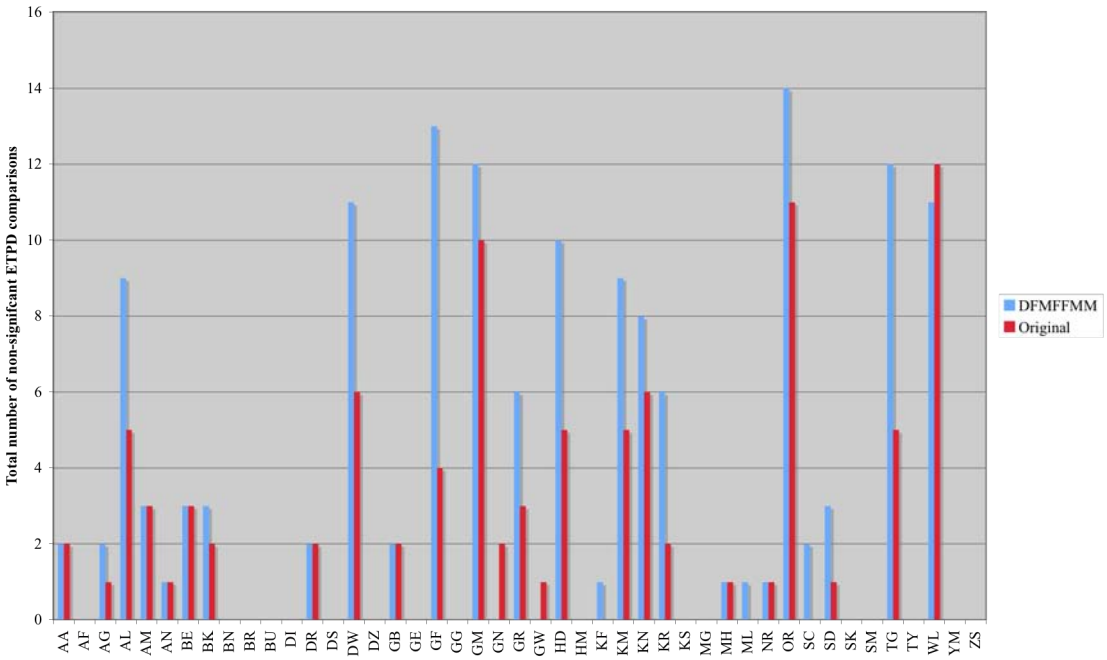
From Figure 4.10, it can be seen that there are 25 incidences where the total number of significant ETPD comparisons is greater in the data restricted to samples with a uniform across generation ethnic identity compared to the original data, and 8 incidences where the total number of significant ETPD comparisons is greater in the original dataset. The greatest increase in the number of non-significant ETPD comparisons was observed in the Kore (KR), which increased from 12 to 18. After removal of across generation ethnically diverse samples, the Kore were no longer distinct from the Bena (BE), Busa (BU), Ganjule (GG), Mashile (MH), Maale (ML) and Sheko (SK). The greatest decrease in the number of non-significant ETPD comparisons was seen in Konta (KN), which decreased from 7 to 4 non-significant differences. After removal of across generation ethnically diverse samples, the Konta were observed to be distinct from the Bench (BN), Dawuro (DW) and Gamo (GM).

Figure 4.11 Total number of non-significant (p>0.01) ETPD comparisons in the original data and across generation ethnicity matched dataset (DFMFFMM), based on the frequency of NRY UEP-MS haplotypes in ethnic groups.



In Figure 4.11, there are 7 cases where the total number of significant ETPD comparisons is greater in the data filtered for samples with a uniform across generation ethnic identity, than in the original data, and a single instance where the total number of significant ETPD comparisons is greater in the original dataset. The greatest increase in the total number of non-significant ETPD comparisons occurred in the Wolayta (WL), which was observed to increase from 1 to 5 non-significant comparisons. After removal of samples of diverse across generation ethnic identity, the Wolayta were no longer distinct from the Gofa (GF), Gamo (GM), Gurage (GR) and Tigray (TG). The only decrease in the number of significant comparisons was observed in the Alaba (AL), which was found to be distinct from the Gurage and Kembata (KM) after removal of samples with a non-uniform across generation ethnic identity.

Figure 4.12 Total number of non-significant (p>0.01) ETPD comparisons in the original data and across generation ethnicity matched dataset (DFMFFMM), based on the frequency of mtDNA HVS1 haplotypes in ethnic groups.



From Figure 4.12, it can be seen that there were 17 instances where the total number of non-significant ETPD comparisons increased after the removal of samples with diverse across generation ethnic identity, and 3 cases where the total number of non-significant comparisons was greater in the original data. The greatest increase in non-significant comparisons was observed in the Gofa (GF), which increased from 4 to 13 compared after removal of samples. After removal of samples with diverse across generation ethnic identity, the Gofa were found to be no longer distinct from the Alaba (AL), Basketo (BK), Dawuro (DW), Hadiya (HD), Kembata (KM), Kore (KR), Maale (ML), Oromo (OR) and Tigray (TG). The greatest decrease in non-significant comparisons was seen to occur in the Genta (GN), which was found to be distinct from both the Gamo (GM) and Wolayta (WL) after removal of samples with non-uniform across generation ethnic identity and consequently were distinct from all other ethnic groups in Ethiopia based on frequencies of mtDNA HVS1 haplotypes.

Table 4.13 reports the results of ETPD comparison between the DFMFFMM sample set and the set of removed samples due to non-uniform across generation ethnic identity. In most cases the removed samples set is no different to the DFMFFMM sample set. However, significant ($p>0.05$) differences were observed in 4 of 42 comparisons using mtDNA haplotypes, 6 of 42 comparisons using NRY UEP haplogroups and 18 of 42 comparisons using NRY UEP-MS haplotype frequencies.

The general decrease in the degree of distinctiveness in the ethnic groups could be explained if many of the samples with diverse across generation ethnic identity had genealogical descent from individuals not local to the ethnic group from which they were removed, but had origins from further afield. Removal of theses samples could result in a greater similarity between geographically proximate ethnic groups that were previously dissimilar due to the differential recent introgression of samples with distinct haplotype distributions. Table 4.12 shows the results of Mantel tests using geographic distance and genetic distances after the removal of samples with a non-uniform across generation ethnic identity. In comparison with the results in Table 4.8, after removal of samples with a diverse across generation ethnic identity, there is a stronger correlation with geographic distance for both NRY and mtDNA genetic distance when assessed by markers that are less susceptible to recent differential introgression or reproductive success (NRY UEP Fst, NRY MS Rst and mtDNA HVS1 K2P), as would be expected if there is genetic isolation due to geographic distance.

Table 4.12 Summary of Mantel test results comparing geographic distance between ethnic groups and genetic distances after removal of samples with diverse ethnic ancestry ($p<0.01$, $p<0.05$)

Metric	Mantel r	P value
NRY UEP Fst	0.3134	0.0082
NRY UEP-MS Fst	0.0572	0.2905
NRY MS Rst	0.2786	0.0160
mtDNA HVS1 Fst	0.2008	0.0620
mtDNA HVS1 K2P	0.5198	0.0001

Table 4.13 Results of ETPD between DFMFFMM sample set and the sample set removed due to diversity in across generation ethnic identity ($p<0.01$, $p<0.05$)

Ethnic group	DFMFFMM sample set (n_1)	Removed sample set (n_2)	ETPD p value using NRY UEP haplogroup frequencies	ETPD p value using NRY UEP-MS haplotype frequencies	ETPD p value using mtDNA HVS1 haplotype frequencies
AA	105	12	***0.003***	***0.004***	0.279
AF	102	10	0.339	0.621	0.721
AG	218	51	0.858	0.507	0.062
AL	83	27	0.186	***0.004***	0.144
AM	337	59	0.699	0.979	0.716
AN	107	1	NA	NA	NA
BE	117	7	0.085	0.304	0.196
BK	99	14	0.476	0.372	0.454
BN	117	10	***0.003***	**0.035**	0.306
BR	108	11	1.000	0.320	0.176
BU	103	23	0.289	**0.023**	0.335
DI	108	24	**0.011**	***<0.001***	0.302
DR	74	34	0.270	**0.032**	**0.039**
DS	95	10	0.172	**0.034**	**0.018**
DW	102	15	0.297	0.258	0.237
DZ	97	7	1.000	**0.040**	0.180
GB	99	14	0.403	0.279	0.354
GE	120	2	0.317	0.536	0.142
GF	62	49	0.205	0.697	0.827
GG	97	12	0.164	***0.002***	0.209
GM	79	130	0.061	0.146	0.659
GN	94	19	0.126	**0.028**	0.084
GR	113	39	0.722	0.277	0.275
GW	115	2	1.000	0.975	0.905
HD	86	41	0.798	0.332	0.917
HM	105	7	0.145	0.207	0.053
KF	103	17	**0.013**	0.096	**0.034**
KM	83	34	0.462	0.051	0.723
KN	96	11	0.298	**0.020**	0.824
KR	88	20	0.142	**0.044**	0.653
KS	98	22	**0.011**	***0.001***	***<0.001***
MG	111	4	0.405	***0.004***	0.879
MH	115	15	1.000	0.987	0.103
ML	97	22	0.453	**0.024**	0.176
NR	118		NA	NA	NA
OR	114	35	0.763	0.676	0.453
SC	106	19	0.217	0.648	0.697
SD	113	13	0.112	***0.001***	0.155
SK	113		NA	NA	NA
SM	90	18	***<0.001***	***<0.001***	0.218
TG	35	31	0.679	0.126	0.241
TY	110	4	1.000	0.347	0.055
WL	73	37	0.945	0.125	0.245
YM	104	4	0.224	***0.008***	0.114
ZS	97	14	0.638	0.102	0.194

4.5 How does the diversity and distinctiveness of an ethnic group in the current generation compare to that ethnic group sampled in previous generations?

Due to the existence of substantial diversity in ethnic ancestry, as indicated by the ethnographic data (Supplementary Table Ethnic), it is possible to investigate to what extent ethnic groups sampled differ in their degree of diversity and distinctiveness if analysis uses the identity of either a parent or grandparent. The NRY and mtDNA genetic data were grouped according to the ethnic affiliations of the sample donor's parents, paternal grandfather and maternal grandmother. It was observed that 5743 donor's father's, 5737 donor's mother's, 5739 donor's paternal grandfather's and 5732 donor's maternal grandmother's were members of one of the 45 ethnic groups listed for the sample donor's generation.

When comparing diversity values for the 45 ethnic groups sampled in different generations, it was observed that the rank order of increasing diversity values was highly correlated ($P<0.001$), although not identical, between generations (see Table 4.16 to Table 4.20). For all measures of diversity, the strongest correlation was observed between the rank order diversity values for ethnic groups sampled in the donor's parents and grandparents generations, whereas the weakest correlation (although still highly significant) was observed between rank order diversity values for ethnic groups sampled in the donors and donor's grandparents generations. For most diversity metrics, the overall change in diversity values between generations did not show a significant deviation from an equal likelihood of an increase or decrease in value (see Supplementary Table DivGens), with the only exception being UEP-MS gene diversity, where it was observed that significantly more ethnic groups sampled in the donor's father's and paternal grandfather's generations had higher diversity values than those ethnic groups sampled in the donor's generation (2-tailed sign test $p<0.01$ for both).

Table 4.14 AMOVA results for NRY data in different generations (for all values p<0.001)

	Donor	Father	Paternal grandfather
UEP Fst	0.12973	0.13390	0.13500
UEP-MS Fst	0.05825	0.06050	0.06133
MS Rst	0.12410	0.12780	0.12906

Table 4.15 AMOVA results for mtDNA data in different generations (for all values p<0.001)

	Donor	Mother	Maternal grandmother
mtDNA Fst	0.00971	0.00968	0.00963
mtDNA K2P	0.02123	0.02161	0.02174

The difference in pairwise genetic distance values (see Supplementary Table DistGens) between ethnic groups sampled in different generations was slight, and the overall pattern of similarity between the ethnic groups did not change substantially compared to that observed for the donor's generation when the donor's parent's and grandparent's generations were sampled. AMOVA (see Table 4.14 and Table 4.15) revealed that there appears to be a slight increase in most global distance values with each preceding generation, however, the opposite slight trend was observed when AMOVA was performed using mtDNA Fst.

Table 4.16 Correlations between ranked NRY UEP gene diversity values in different generations (p-values in upper diagonal, correlation coefficients in lower diagonal)

	Donor	Father	Paternal grandfather
Donor	*	<0.001	<0.001
Father	0.9930	*	<0.001
Paternal grandfather	0.9813	0.9937	*

Table 4.17 Correlations between ranked NRY UEP-MS gene diversity values in different generations (p-values in upper diagonal, correlation coefficients in lower diagonal)

	Donor	Father	Paternal grandfather
Donor	*	<0.001	<0.001
Father	0.9866	*	<0.001
Paternal grandfather	0.9743	0.9921	*

Table 4.18 Correlations between ranked NRY MSV values in different generations (p-values in upper diagonal, correlation coefficients in lower diagonal)

	Donor	Father	Paternal grandfather
Donor	*	<0.001	<0.001
Father	0.9931	*	<0.001
Paternal grandfather	0.9915	0.9966	*

Table 4.19 Correlations between ranked mtDNA HVS1 gene diversity values in different generations (p-values in upper diagonal, correlation coefficients in lower diagonal)

	Donor	Mother	Maternal grandmother
Donor	*	<0.001	<0.001
Mother	0.9772	*	<0.001
Maternal grandmother	0.9695	0.9909	*

Table 4.20 Correlations between ranked mtDNA HVS1 nucleotide diversity values in different generations (p-values in upper diagonal, correlation coefficients in lower diagonal)

	Donor	Mother	Maternal grandmother
Donor	*	<0.001	<0.001
Mother	0.9756	*	<0.001
Maternal grandmother	0.9665	0.9920	*

ETPD (see Supplementary Table ETPD and Supplementary Table ETPDGens) using NRY haplotypes, it was observed that there was a general decrease in the number of non-significant ETPD comparisons with each preceding generation, with a total of 26 non-significant differences observed in the Donors generation, 14 in the father's generation and 10 in the paternal grandfather's generation (Figure 4.14). The largest change was in the Tigray (TG), which was observed to be significantly differentiated from the Alaba (AL), Gurage (GR), Kembata (KM) and Oromo (OR) when sampled with respect to the father's generation, and also the Amhara (AM) when sampled with respect to the paternal grandfather's generation, but was observed to be non-distinct from the above in the donor's generation (Figure 3.13, Figure 4.16, Figure 4.17).

For ETPD using mtDNA haplotypes it was observed that overall there were more non-significant differences in the preceding generations (128 in the mother's, and 118 in the maternal grandmother's generation) than in the donor's generation (96 overall non-

significant differences). The most marked change was in the Gamo (GM) which was observed to be far more distinct in the donor's generation (non-differentiated from 10 ethnic groups) than in the donor's mother's or maternal grandmother's generation (non-differentiated from 20 and 16 ethnic groups respectively, see Figure 4.15). Interestingly, the Wolayta (WL), which were observed to be the least distinct ethnic group when sampled in the donor's generation, were however observed in the mother's generation to be distinct from the Alaba (AL), Basketo (BK), Genta (GN), Hadiya (HD), Kore (KR), and Tigray (TG), and also the Gofa (GF), Kembata (KM) and Konta (KN) (although not Hadiya) in the maternal grandmother's generation, but non-distinct from the above in the donor's generation (see Supplementary Table ETPD, and Supplementary Table ETPDGens).

The above patterns of variation amongst ethnic groups over generations in the NRY and mtDNA genetic systems are consistent with a greater introgression of haplotypes into ethnic groups from a non-local origin in the current generation compared with that in preceding two generations. This would also be consistent with 'adopted' ethnicities in the current generation (ethnic identity of the donor differs from both parents) being greater than in previous generations. The observed pattern of NRY Fst being progressively lower in each generation from the grandparent to the donor generation while the mtDNA showed a small increase is consistent with male movement between ethnic groups, introgression from a common external source into the different groups in relatively similar proportions to their sizes, or both, while female movement between the groups and introgression from outside them was, at the least, very restricted.

Figure 4.13 Total number of non-significant (p>0.01) ETPD comparisons in different generations, based on the frequency of NRY UEP haplogroups in ethnic groups.

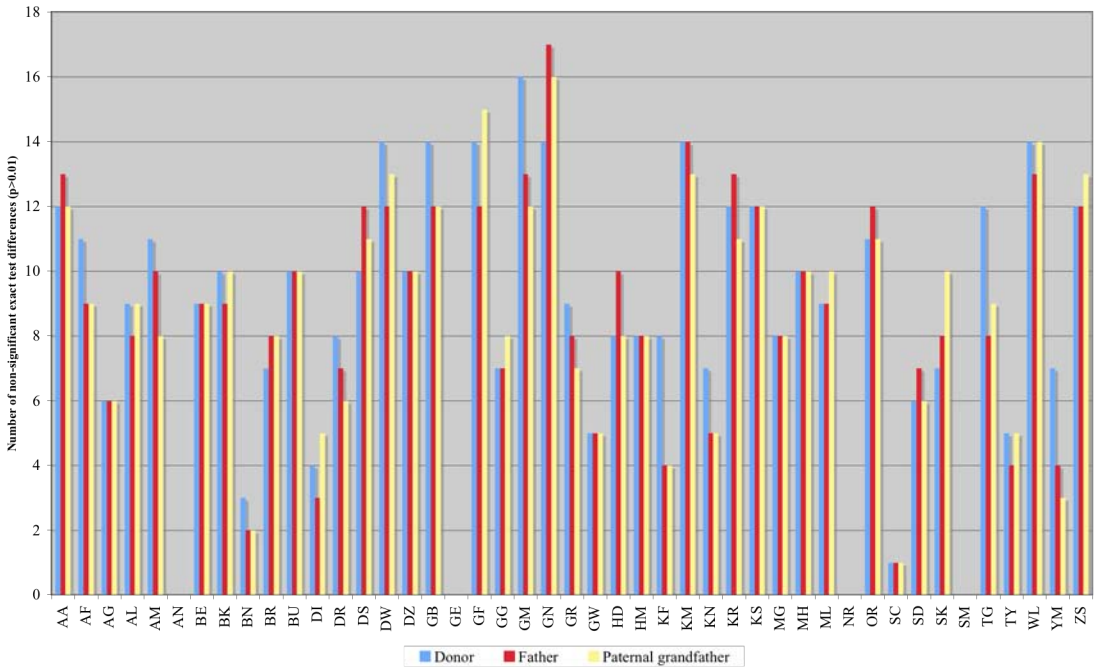


Figure 4.14 Total number of non-significant (p>0.01) ETPD comparisons in different generations, based on the frequency of NRY UEP-MS haplotypes in ethnic groups.

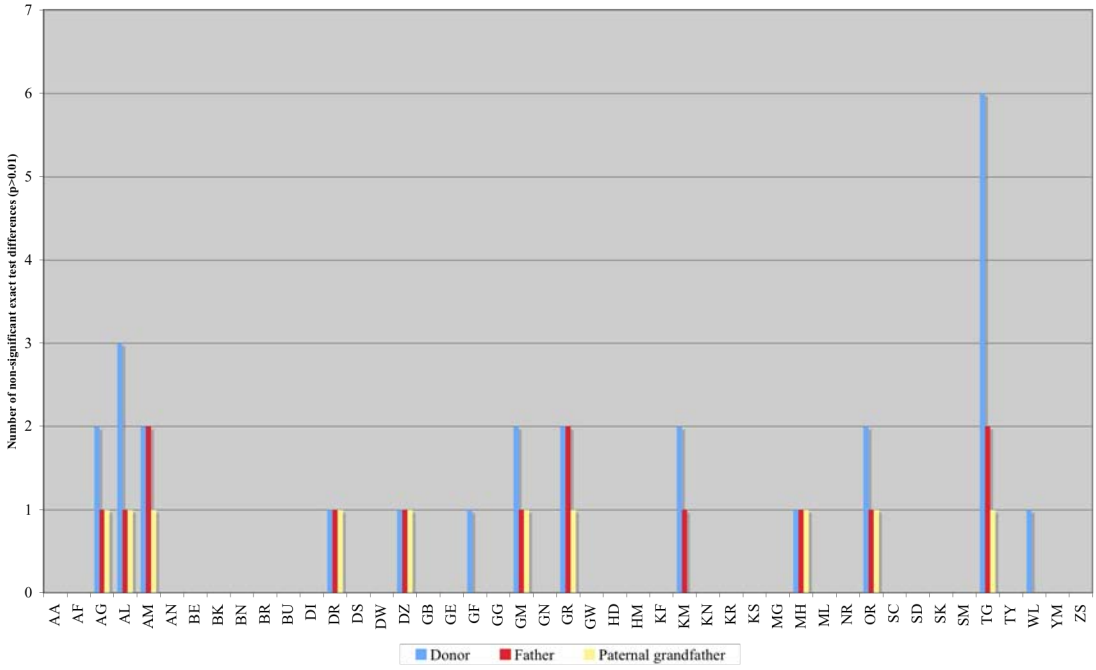


Figure 4.15 Total number of non-significant (p>0.01) ETPD comparisons in different generations, based on the frequency of mtDNA HVS1 haplotypes in ethnic groups.

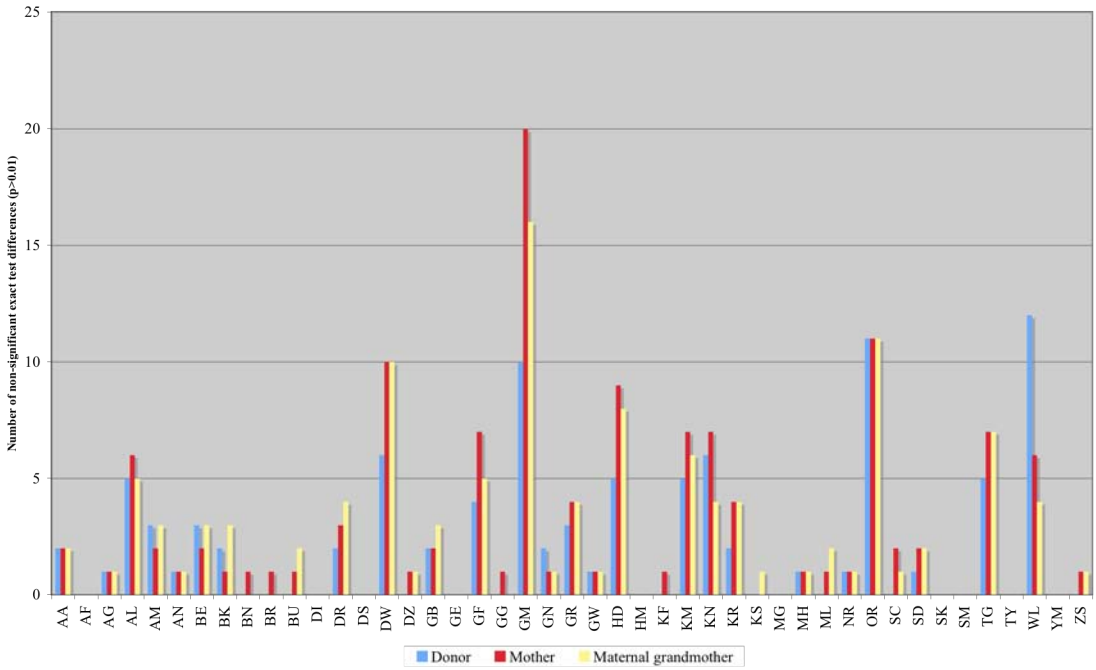


Figure 4.16 Maps showing non-significant (p>0.01) ETPD comparisons (indicated by red lines) between Ethiopian ethnic groups (indicated by dots) sampled in the father's generation, using NRY UEP-MS haplotype frequencies.

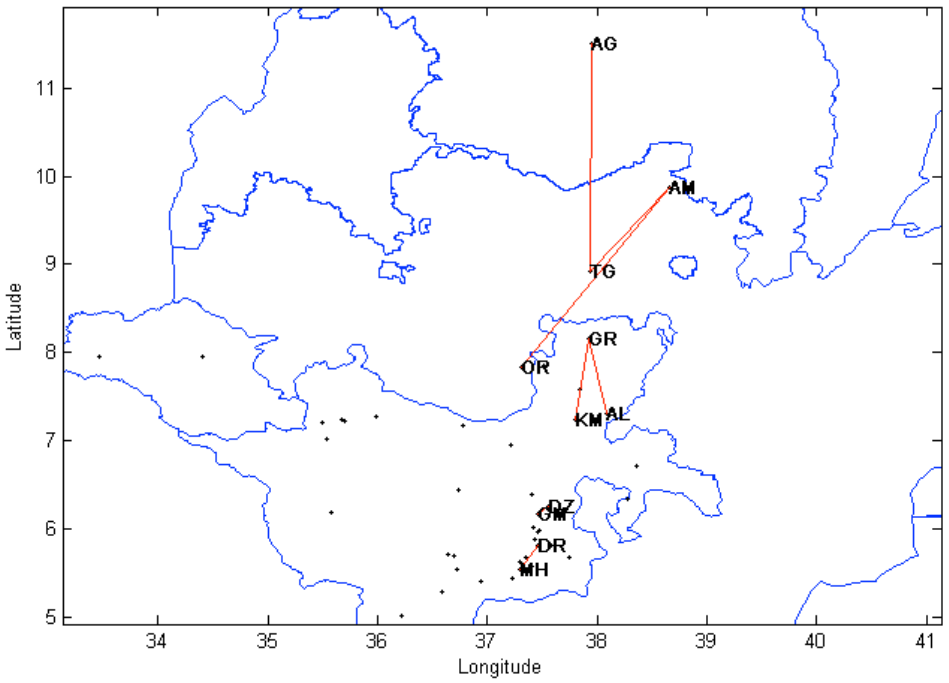


Figure 4.17 Maps showing non-significant (p>0.01) ETPD comparisons (indicated by red lines) between Ethiopian ethnic groups (indicated by dots) sampled in the paternal grandfather's generation, using NRY UEP-MS haplotype frequencies.

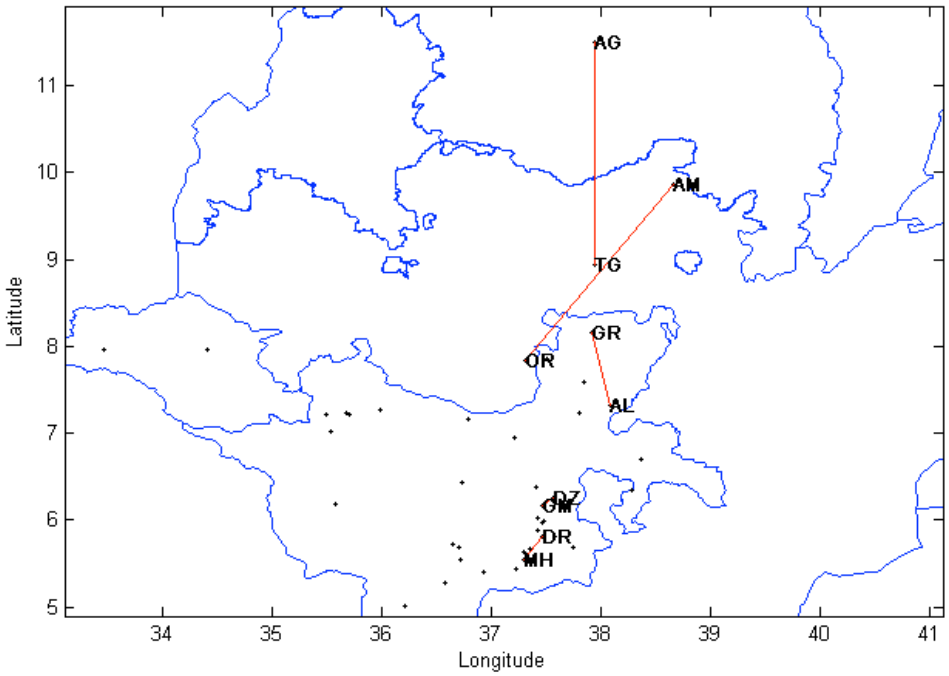


Figure 4.18 Maps showing non-significant (p>0.01) ETPD comparisons (indicated by red lines) between Ethiopian ethnic groups (indicated by dots) sampled in the mother's generation, using mtDNA HVS1 haplotype frequencies.

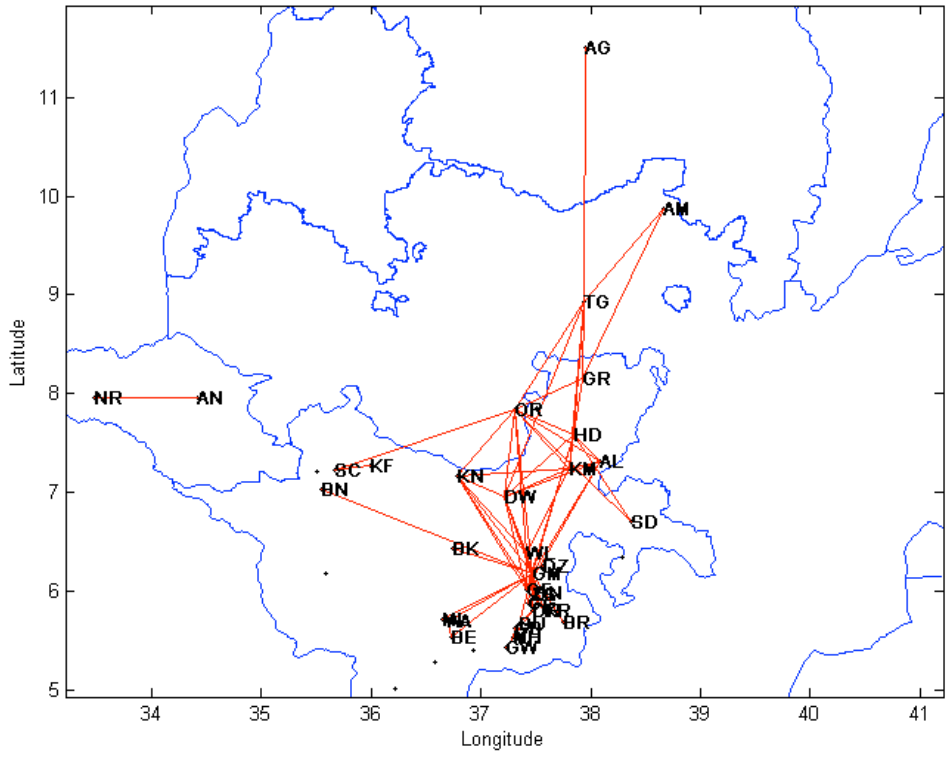
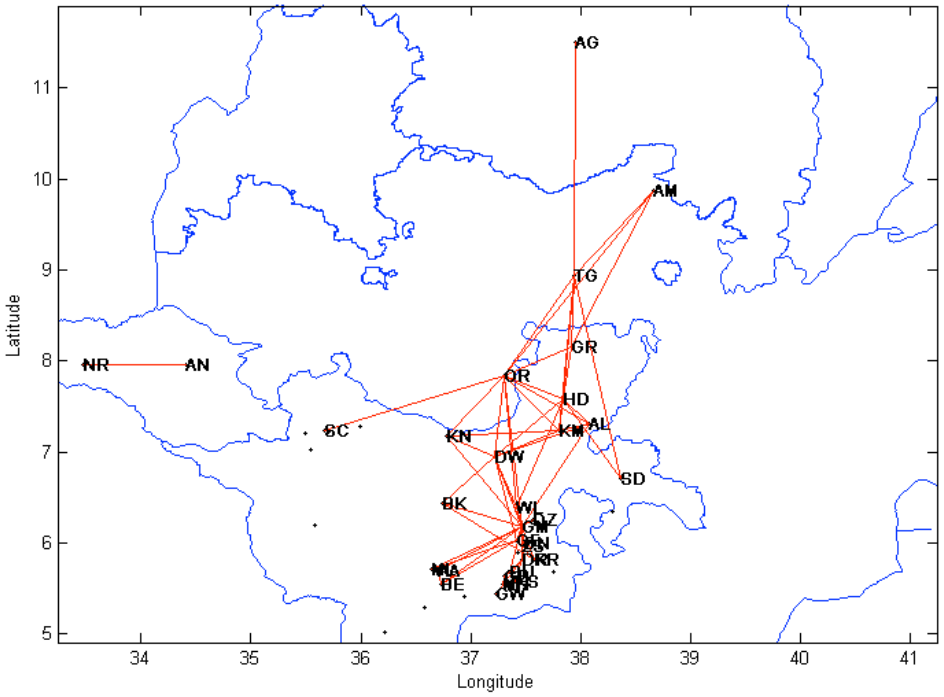


Figure 4.19 Maps showing non-significant (p>0.01) ETPD comparisons (indicated by red lines) between Ethiopian ethnic groups (indicated by dots) sampled in the maternal grandmother's generation, using mtDNA HVS1 haplotype frequencies.



Chapter 5: To what extent does increasing the number of uniparental markers increase the ability to differentiate ethnic groups?

A subset of the total Ethiopian collection (Ethiopian ascertainment samples), was chosen in which additional uni-parental markers were genotyped. This subset comprised of up to 74 samples of the Afar (AF), 76 of the Amhara (AM), 76 of the Oromo (OR), 75 of the Maale (ML) and 76 of the Anuak (AN) ethnic groups, representing each of the four linguistic groups in Ethiopia in a rough geographical north-east to south-west transect.

5.1 Does increasing the number of NRY STR markers lead to an increase in the power to discriminate between ethnic groups and alter the pattern of their relationships?

14 NRY STR loci (DYS19, DYS390, DYS391, DYS392, DYS393, DYS389I, DYS389II, DYS437, DYS438, DYS439, DYS448, DYS456, DYS635, Y GATA H4) of the AmpFℓSTR Yfiler (Applied Biosystems) were successfully assayed in 331 samples, and the data combined with the results of DYS388 locus from the MS1 kit (Thomas et al. 1999) for these samples. As five of these loci (DYS19, DYS390, DYS391, DYS392, DYS393) overlap with those assayed by the MS1 kit, it was possible to estimate a potential genotyping error rate. It was observed that for 10 samples, the STR repeat sizes using the AmpFℓSTR Yfiler kit for the five overlapping loci did not all match the sizes determined by the MS1 assay for same loci (3% error rate), with 8 samples with a single locus differing from that determined by the MS1 kit, and 2 samples with two or more differing loci. There could be several potential causes for this discrepancy between the in-house and outsourced results, including human error in reading or entering the data, and sample-to-sample cross contamination, either during transit, or in the laboratory (which can be a particular risk when working in an environment with both genomic template and amplified DNA). Consequently, all resultant analysis was performed on only the 321 samples with unambiguous data (see Supplementary Table NRY15STR).

Table 5.1 shows the variance values for each of the 15 STRs assayed, the MSV for each of the 5 ethnic groups, and the overall values for all the data combined. The highest 15 STR MSV was observed in the Amhara (0.942, AM), whereas the lowest was observed in the Anuak (0.617, AN), with an overall MSV for the 321 samples of 0.936. Using the original 6 STRs, the highest MSV was also observed in the Amhara (1.296) and the lowest was observed in the Maale (0.673, ML), with an overall MSV for 6 STRs of 1.234. The rank order of increasing MSV for the ethnic groups using 6 STRs and 15 STRs was similar, although not significant (Spearman's p=0.083), with only the Anuak exchanging places with the Maale for position as the group with the lowest MSV with the addition of the 9 STRs. With the exception of the Maale, the inclusion of the additional 9 STRs resulted in a substantial decrease in MSV values.

Table 5.2 shows the gene diversity values estimated from the frequency of 15 STR and 6 STR haplotypes in the five ethnic groups. Increasing the number of STR markers from 6 to 15 increased the number of haplotypes resolved in ethnic groups by varying degrees, and also increased the number of private haplotypes (haplotypes restricted to a single ethnic group, as depicted in the median joining networks in Figure 5.1 to Figure 5.4). The largest increase in resolution was observed in the Amhara, which doubled the number of observed haplotypes from 34 to 68. However, increasing the number of STR markers had the least effect on resolving additional haplotypes in the Maale, which only increased from 26 to 31 (19.2% increase). Interestingly, the number of haplotypes resolved in the other three ethnic groups (Afar, Anuak and Oromo) after the inclusion of 9 additional STRs all increased by very similar amounts (63.3%, 64.0% and 58.3% respectively). Increasing the number STR markers from 6 to 15 resulted in the mean number of resolved private haplotypes increasing from 52.3% to 91.7% of haplotypes, with 100% of the Anuak haplotypes only observed in the Anuak. The rank order of increasing *h* values was the same when using both the 15 STR and the 6 STR haplotypes (Spearman's p=0.017), with highest gene diversity observed in the Amhara, and lowest observed in the Anuak.

Figure 5.5 and Figure 5.6 show PCO plots of the pairwise Rst distances between five ethnic groups using 15 NRY STRs and 6 NRY STRs respectively (Table 5.3 and Table 5.4 respectively), whereas Figure 5.7 and Figure 5.8 show PCO plots of the pairwise Fst

distances between five ethnic groups using 15 NRY STRs and 6 NRY STRs respectively (Table 5.5 and Table 5.6 respectively) The Anuak are the clear outliers along PCO 1 in all plots, with all other groups spreading out along PCO1 in the Rst plots, and along PCO2 in the Fst plots. However, the most striking feature appears to be that the use of 15 STRs does not seem to substantially change the relative positions of the ethnic groups to each other compared to the pattern observed for 6 STRs, and for both Rst plots and for both Fst plots the proportion of the variance explained by PCO1 is very similar.

Table 5.1 Variance values for 15 STRs in five ethnic groups

STR	DYS19	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS635	Y GATA H4	15 STR MSV	6 STR MSV	n
AF	1.160	2.521	0.584	1.311	1.673	0.254	0.405	0.397	0.498	0.080	1.252	0.641	0.652	0.949	0.279	0.844	1.068	63
AM	1.111	3.534	0.480	1.001	2.005	0.473	0.182	0.469	0.547	0.360	0.775	0.478	0.650	1.662	0.404	0.942	1.296	70
AN	1.143	0.677	0.240	0.827	1.483	0.391	0.046	0.372	0.447	0.297	0.691	0.770	0.281	1.078	0.512	0.617	0.685	63
ML	0.608	1.170	0.465	1.745	1.102	0.570	0.181	0.408	0.086	0.209	0.491	1.256	1.254	1.524	0.117	0.746	0.673	57
OR	2.517	2.615	0.187	1.100	1.239	0.313	0.299	0.512	0.479	0.310	0.610	0.615	1.142	1.119	0.427	0.899	1.249	68
Overall	1.887	2.329	0.407	1.245	2.029	0.412	0.234	0.516	0.434	0.323	0.862	0.872	0.817	1.291	0.381	0.936	1.234	321

Table 5.2 Gene diversity values and numbers of haplotypes for 15 STR and 6 STR datasets in five ethnic groups

Ethnic group	Total 15 STR haplotypes	Proportion of private 15 STR haplotypes	15 STR *h*	s.d.	Total 6 STR haplotypes	Proportion of private 6 STR haplotypes	6 STR *h*	s.d.
AF	49	0.918	0.990	0.005	30	0.533	0.952	0.012
AM	68	0.882	0.999	0.003	34	0.382	0.965	0.010
AN	41	1.000	0.958	0.018	25	0.680	0.886	0.031
ML	31	0.871	0.975	0.008	26	0.577	0.952	0.014
OR	57	0.912	0.971	0.015	36	0.444	0.929	0.024

Figure 5.1 Median joining network of haplotypes using 6 NRY STRs shaded by ethnic group

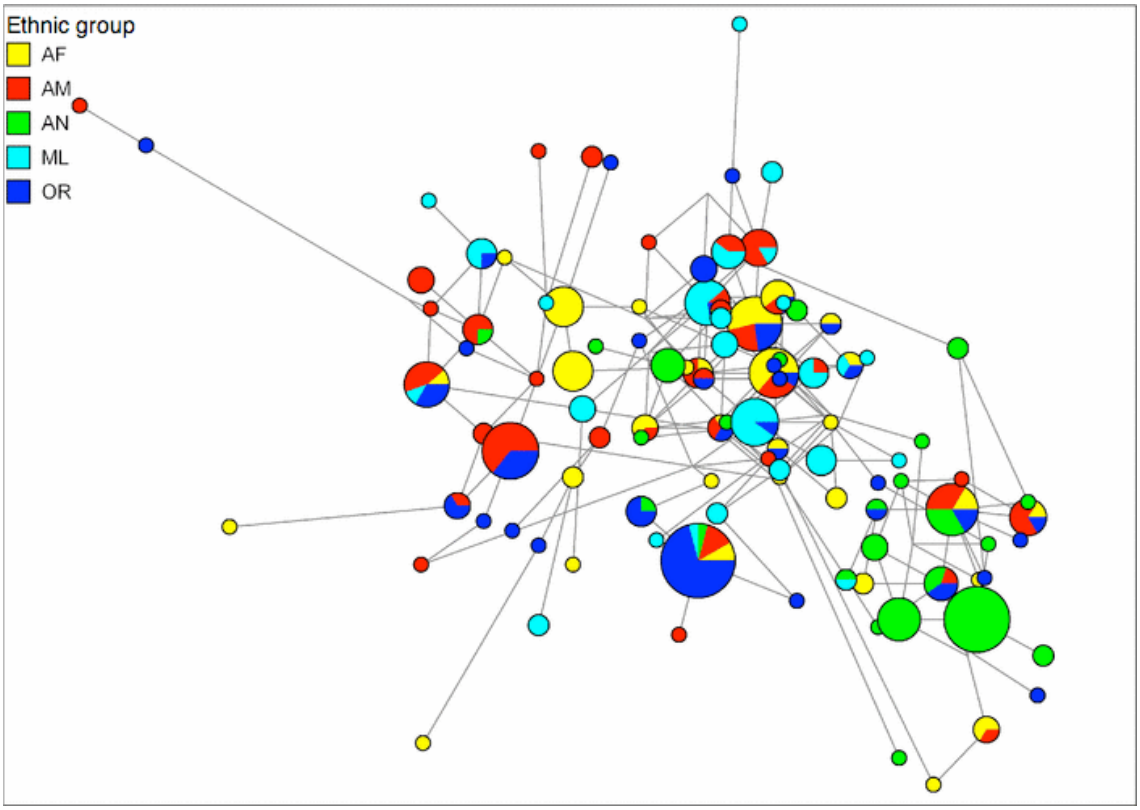


Figure 5.2 Median joining network of haplotypes using 6 NRY STRs shaded by haplogroup

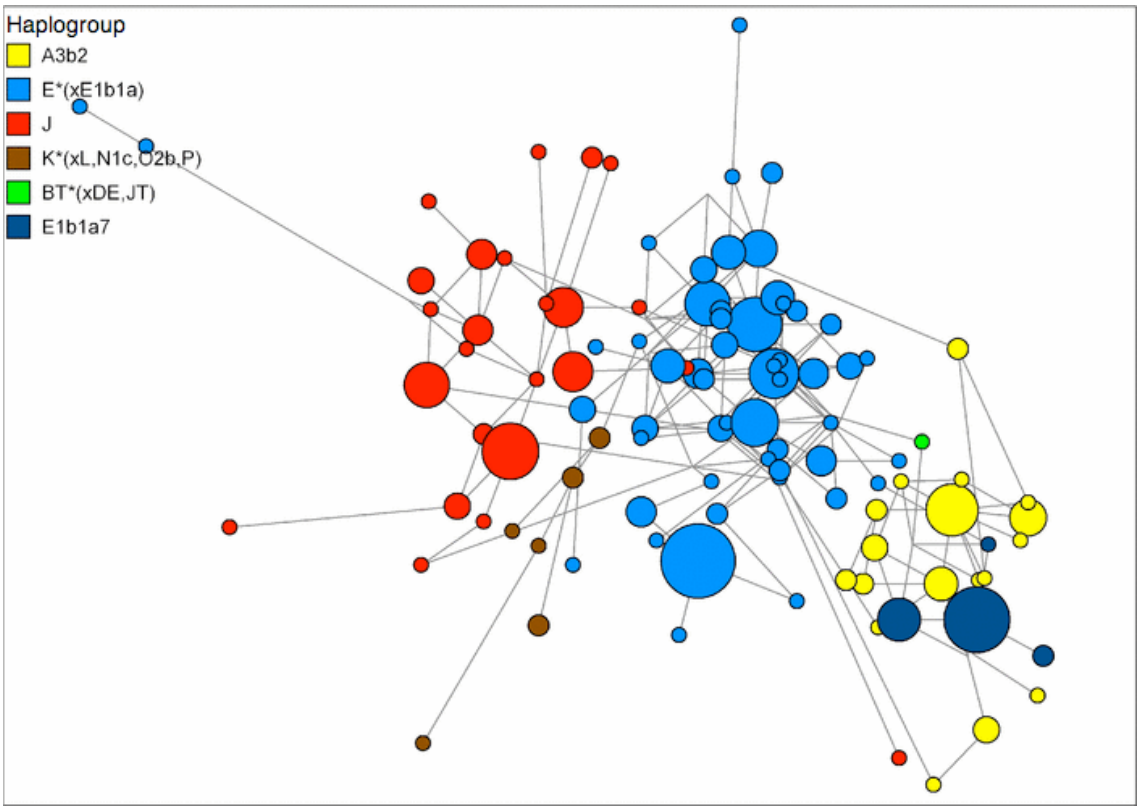


Figure 5.3 Median joining network of haplotypes using 15 NRY STRs shaded by ethnic group

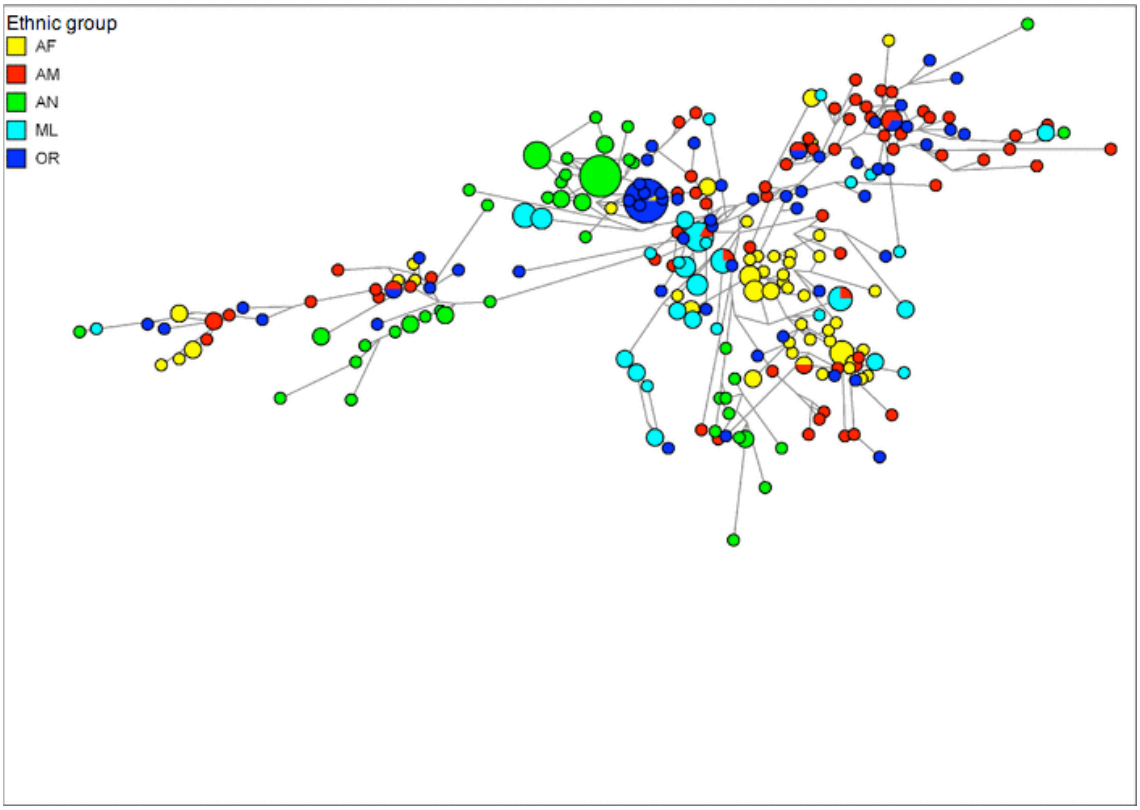


Figure 5.4 Median joining network of haplotypes using 15 NRY STRs shaded by haplogroup

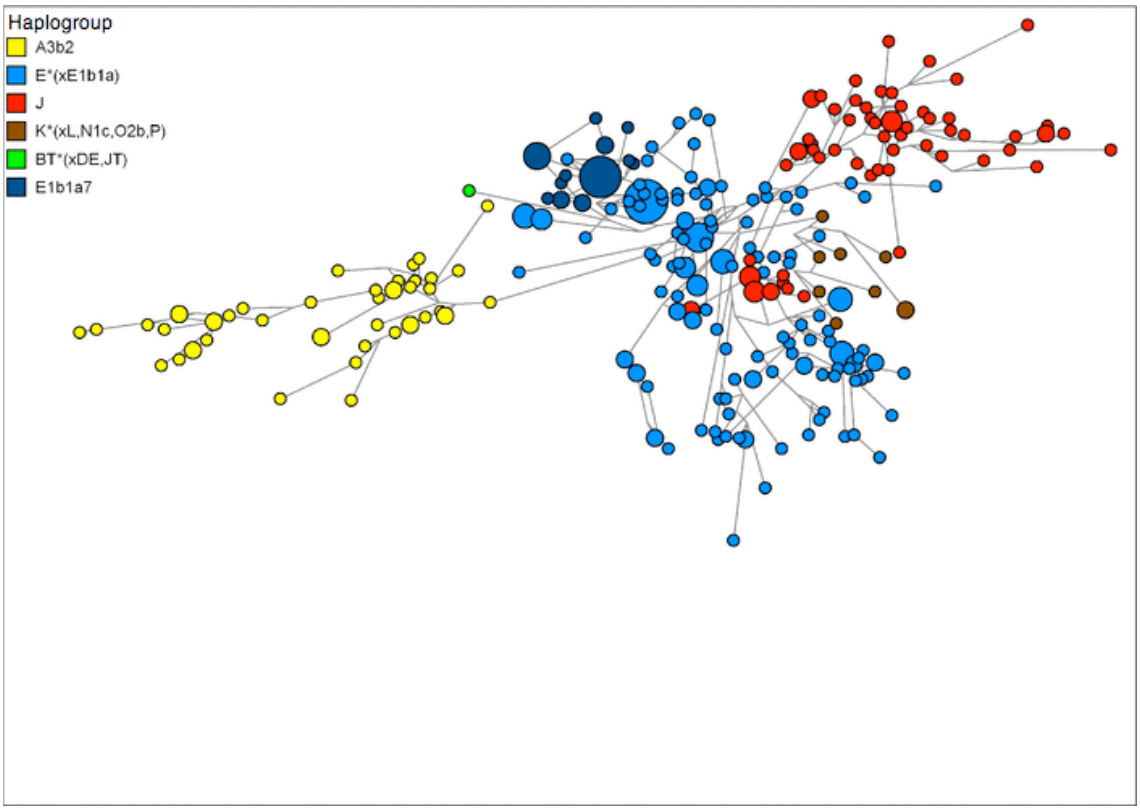


Table 5.3 Pairwise Rst distances (lower diagonal) and p values (upper diagonal) between five ethnic groups using 15 NRY STRs

	AF	AM	AN	ML	OR
AF	*	0.005	<0.001	<0.001	<0.001
AM	0.038	*	<0.001	<0.001	0.002
AN	0.268	0.307	*	<0.001	<0.001
ML	0.089	0.090	0.409	*	<0.001
OR	0.063	0.045	0.357	0.070	*

Table 5.4 Pairwise Rst distances (lower diagonal) and p values (upper diagonal) between five ethnic groups using 6 NRY STRs

	AF	AM	AN	ML	OR
AF	*	0.011	<0.001	<0.001	0.001
AM	0.044	*	<0.001	<0.001	0.001
AN	0.341	0.446	*	<0.001	<0.001
ML	0.084	0.112	0.658	*	0.002
OR	0.077	0.066	0.571	0.065	*

Table 5.5 Pairwise Fst distances (lower diagonal) and p values (upper diagonal) between five ethnic groups using 15 NRY STRs

	AF	AM	AN	ML	OR
AF	*	<0.001	<0.001	<0.001	<0.001
AM	0.005	*	<0.001	<0.001	<0.001
AN	0.026	0.022	*	<0.001	<0.001
ML	0.017	0.010	0.034	*	<0.001
OR	0.017	0.014	0.036	0.027	*

Table 5.6 Pairwise Fst distances (lower diagonal) and p values (upper diagonal) between five ethnic groups using 6 NRY STRs

	AF	AM	AN	ML	OR
AF	*	<0.001	<0.001	<0.001	<0.001
AM	0.025	*	<0.001	<0.001	<0.001
AN	0.082	0.071	*	<0.001	<0.001
ML	0.048	0.035	0.084	*	<0.001
OR	0.043	0.024	0.089	0.050	*

Figure 5.5 PCO of pairwise Rst distances between five Ethiopian ethnic groups using 15 NRY STRs

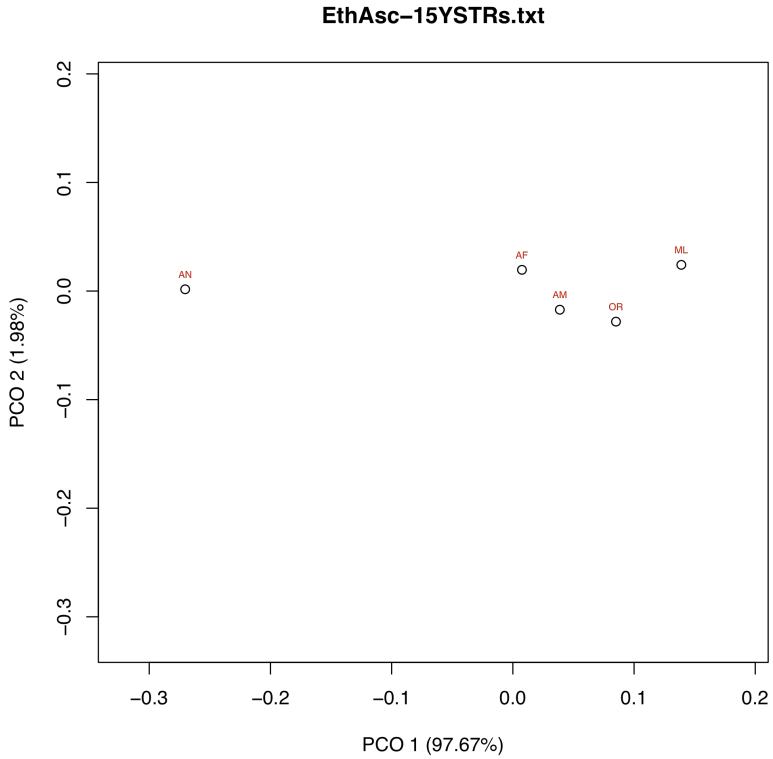


Figure 5.6 PCO of pairwise Rst distances between five Ethiopian ethnic groups using 6 NRY STRs

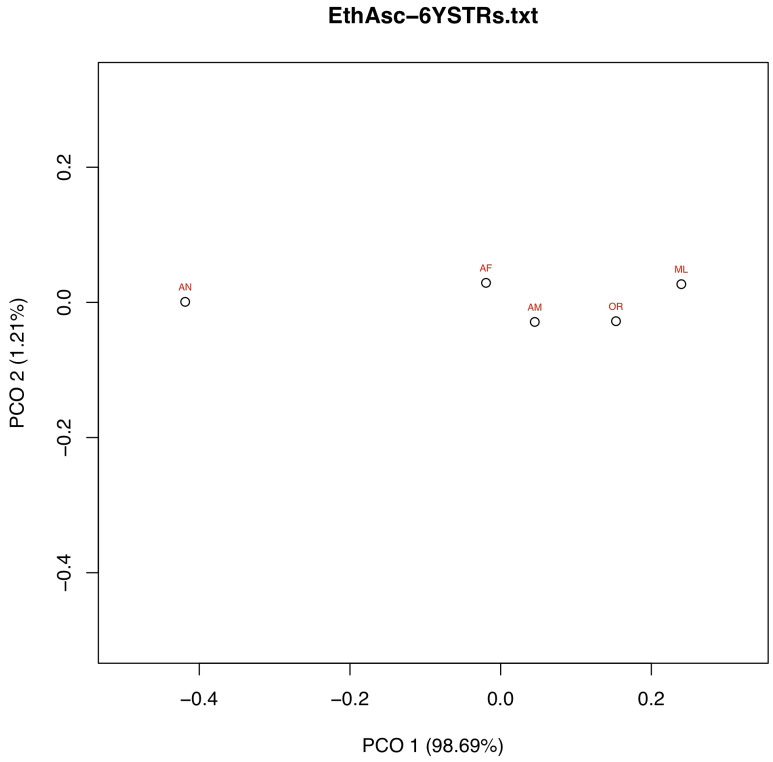


Figure 5.7 PCO of pairwise Fst distances between five Ethiopian ethnic groups using 15 NRY STRs

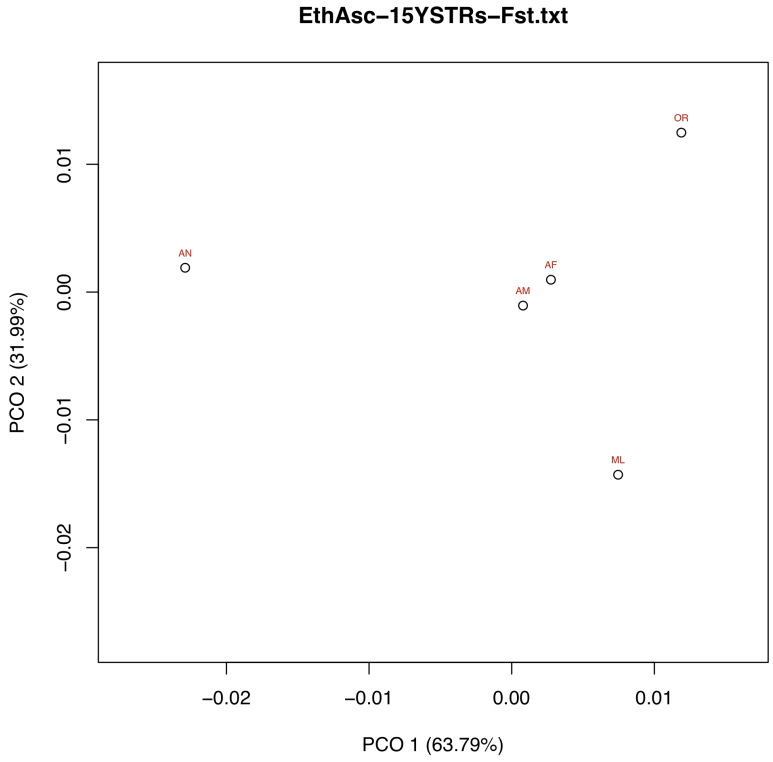
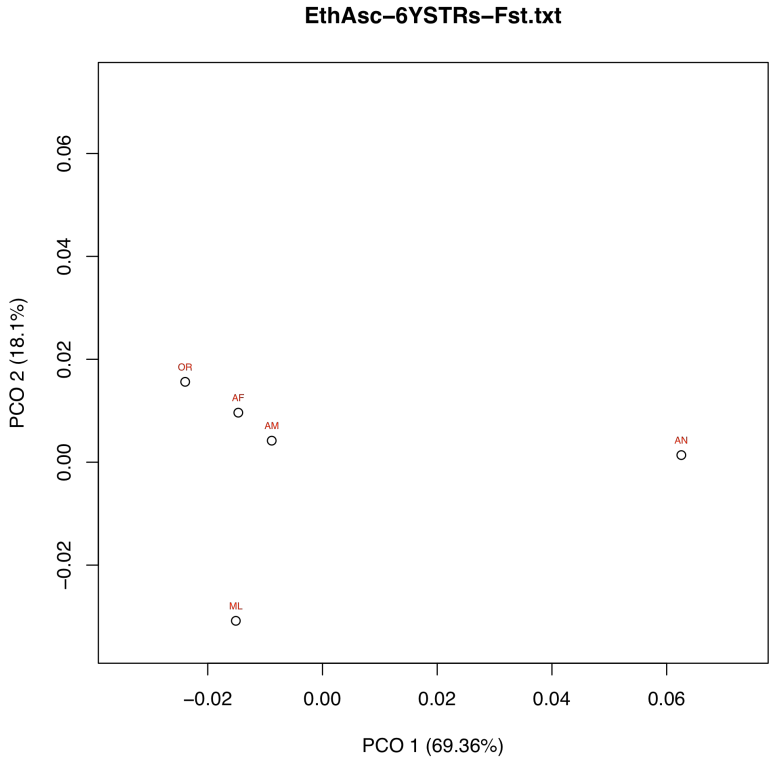


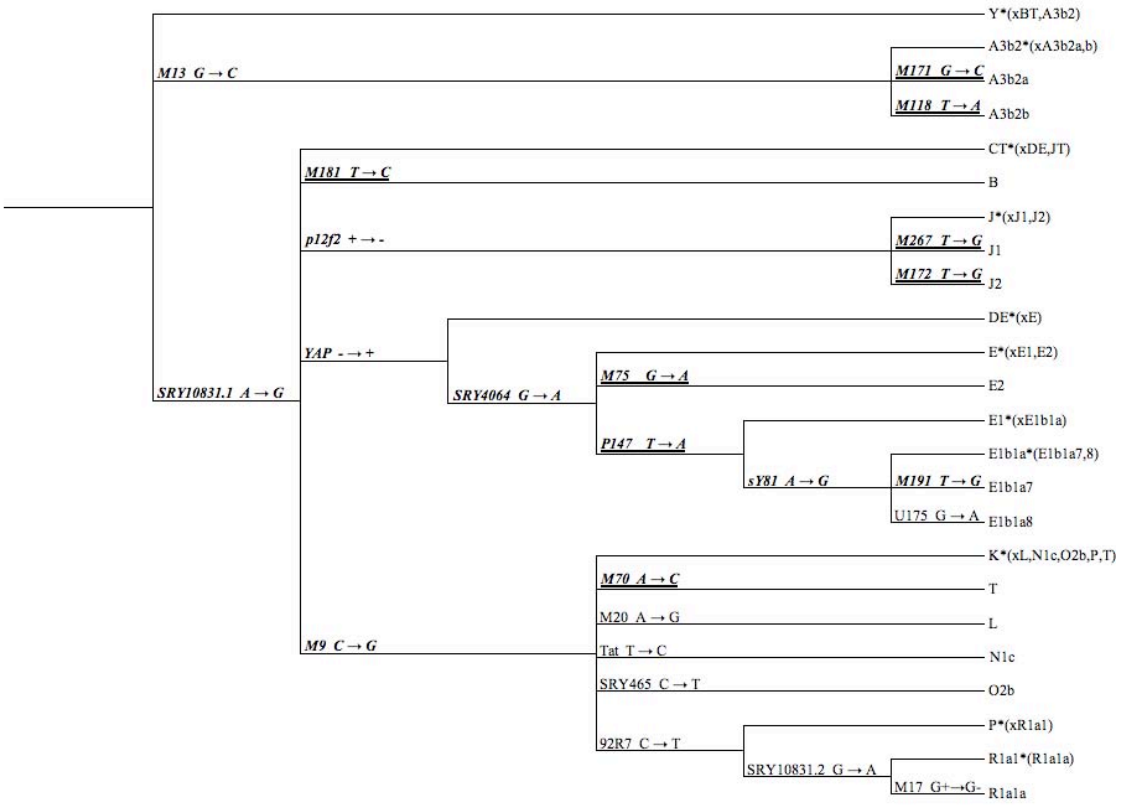
Figure 5.8 PCO of pairwise Fst distances between five Ethiopian ethnic groups using 6 NRY STRs



5.2 Higher resolution Y chromosome markers and the dating of clades present in Ethiopia

Five ethnic groups were genotyped for additional NRY markers (see Chapter 2) to further resolve the haplogroups present in these groups (Figure 5.9).

Figure 5.9 Genealogical relationship of haplogroups defined by UEP and the additional NRY haplogroup markers according to the nomenclature of Karafet et al. (2008). Markers with derived states of UEP markers in the Ethiopian ascertainment samples are indicated in italic type, additional haplogroup markers genotyped are in italic and underlined



Complete results of the genotyping, including the partially resolved haplogroups are shown in Supplementary Table NRYaddHG. Table 5.7 summarises the results of the 349 samples with fully resolved haplogroups resulting from the genotyping in the five ethnic groups. Genotyping of two markers in A3b2 (defining haplogroups A3b2a and A3b2b) clade revealed that nearly all A3b2 samples are A3b2b, with highest frequency in the Anuak (AN, 22%), and was not observed in the Maale (ML). Samples that did not belong to either clade (A3b2*(xA3b2ab)) were found at a low frequency in the Anuak (3% frequency) and the Maale (1% frequency). All samples that were of

haplogroup BT*(xDE,JT), where genotyping was successful, were found to be of the B clade, and was only observed in the Anuak (14%). All samples that were of haplogroup K*(xL,N1c,O2b,P), where genotyping was successful, belonged to haplogroup T, and were found at highest frequency (3%) in the Amhara (AM) and Afar (AF). Genotyping for markers defining haplogroups J1 and J2 in the J clade found that the majority of samples belonged to haplogroup J1, with the highest frequencies of both J1 and J2 observed in the Amhara (32% and 4% respectively). A single sample (1% frequency) that did not belong to either of the defined J haplogroups J*(xJ1,J2) was only observed in the Maale. Samples that were of haplogroup E*(xE1b1a), were genotyped for markers defining haplogroups E1 and E2, and it was observed that in the Afar, Amhara, Maale and Oromo (OR), nearly all samples belonged to haplogroup E1, with a low frequency (3%) of samples in the Amhara that did not belong to either of the defined clades (E*(xE1,E2)). Interestingly, almost all E*(xE1b1a) samples in the Anuak were observed to belong to the E2 clade (15%), with only a single sample (1%) belonging to the E1 clade. Note that all frequencies reported for samples belonging to the E1 clade exclude any samples belonging to the more derived E1b1a7 clade, which is reported separately.

Figure 5.10 Frequency of NRY haplogroups present in five ethnic groups

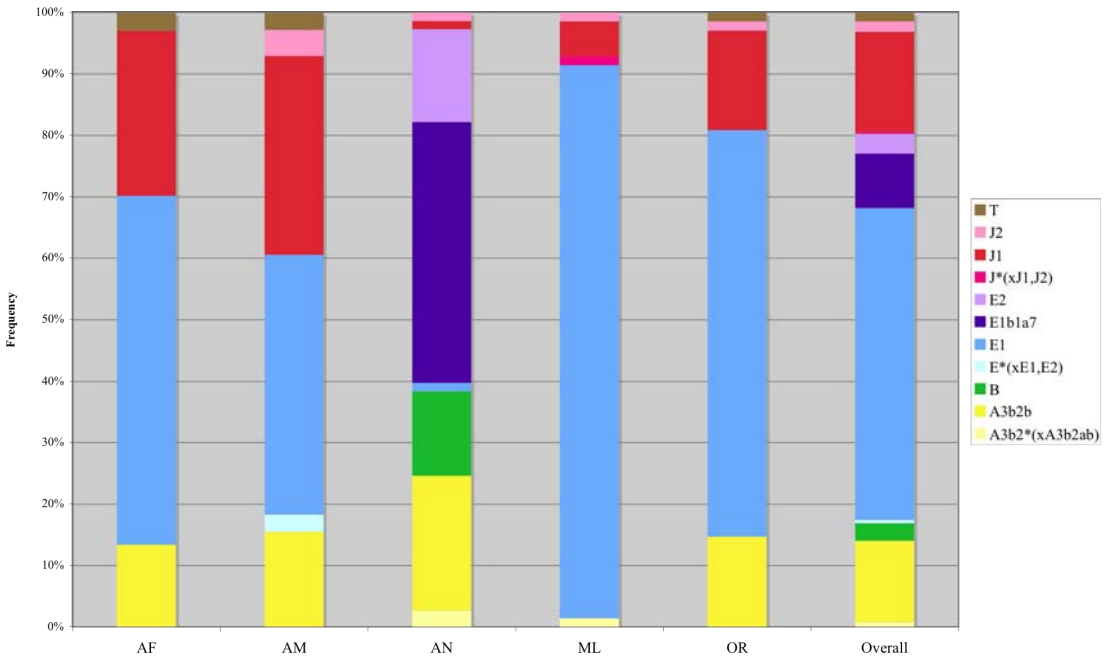


Table 5.7 Frequency of NRY haplogroups present in five ethnic groups (number of samples in parentheses)

Haplogroup	AF	AM	AN	ML	OR	Overall
A3b2*(xA3b2ab)	0.00 (0)	0.00 (0)	0.03 (2)	0.01 (1)	0.00 (0)	0.01 (3)
A3b2b	0.13 (9)	0.15 (11)	0.22 (16)	0.00 (0)	0.15 (10)	0.13 (46)
B	0.00 (0)	0.00 (0)	0.14 (10)	0.00 (0)	0.00 (0)	0.03 (10)
E*(xE1,E2)	0.00 (0)	0.03 (2)	0.00 (0)	0.00 (0)	0.00 (0)	0.01 (2)
E1	0.57 (38)	0.42 (30)	0.01 (1)	0.90 (63)	0.66 (45)	0.51 (177)
E1b1a7	0.00 (0)	0.00 (0)	0.42 (31)	0.00 (0)	0.00 (0)	0.09 (31)
E2	0.00 (0)	0.00 (0)	0.15 (11)	0.00 (0)	0.00 (0)	0.03 (11)
J*(xJ1,J2)	0.00 (0)	0.00 (0)	0.00 (0)	0.01 (1)	0.00 (0)	<0.01 (1)
J1	0.27 (18)	0.32 (23)	0.01 (1)	0.06 (4)	0.16 (11)	0.16 (57)
J2	0.00 (0)	0.04 (3)	0.01 (1)	0.01 (1)	0.01 (1)	0.02 (6)
T	0.03 (2)	0.03 (2)	0.00 (0)	0.00 (0)	0.01 (1)	0.01 (5)
No. of samples	67	71	73	70	68	349

296 samples were successfully genotyped for both the additional NRY haplogroup markers and the 15 NRY STR markers (Table 5.8, Supplementary Table NRYaddHGSTR). Figure 5.11 and Figure 5.12 indicate the degree of sharing of 15 NRY STR haplotypes and the relative phylogenetic relationship of the haplogroups respectively, in median joining networks. Median joining networks of 15 NRY STR haplotypes were also constructed for each of the four most common haplogroups: A3b2b, E1, E1b1a7 and J1 (Figure 5.13 to Figure 5.16 respectively), with branch lengths proportional to the number of STR repeats separating nodes.

From Figure 5.13, the most striking feature of the relationship amongst the STR haplotypes in A3b2b is that the majority of the haplotypes found in the Anuak (AN) form a cluster away from those found in the other ethnic groups, with many of the haplotypes one-step neighbours of each other. The network of E1 haplotypes (Figure 5.14) shows that all the ethnic groups (with the exception of the Anuak, with only one representative haplotype of this clade), have NRY STR haplotypes that appear in many areas of the network. The highest frequency haplotype in this clade predominantly

occurs in the Oromo (OR), but is also found at low frequency in the Afar (AF). Haplogroup E1b1a7 was only observed in the Anuak, and was the only haplogroup where a clear modal haplotype and a star-like pattern were observed (Figure 5.15). Figure 5.16 shows the network of haplotypes for haplogroup J1, and it is apparent that there are two clusters with many one-step neighbour haplotypes. One cluster appears at the bottom of the network, with haplotypes found predominantly in the Oromo and Amhara (AM). The other one-step neighbour cluster appears in the top right of the network, and is exclusively represented by Afar haplotypes.

Supplementary Table NRYSTRdate shows the age of the observed STR variation in the four most common haplogroups defined using the additional NRY markers, and the three most common haplogroups using the original UEP markers, using two different mutation rates. The difference in the ages of variation determined using these two different rates is substantial, with the point estimate ages of variation using the Zhivotovsky et al. (2004) rate up to approximately four fold higher than that obtained by using the rate from YHRD. Both, however, had wide 95% confidence intervals. This large discordance between rates determined by direct observation using pedigree data and rates determined using population divergence time is well recognised, with Zhivotovsky et al. (2006) suggesting one possible explanation for this difference, in that much of the novel variation in haplotypes in populations that have diverged over time has been lost due to genetic drift, resulting in lower levels of observed diversity than that seen in pedigrees, resulting, in turn, in lower estimates for mutation rates. In any case, the rank order of the ages of variation of haplogroups determined using these two mutation rates is the same (see Supplementary Table NRYSTRdate). The oldest haplogroup, with the largest MSV was E1 (MSV=0.584), and from the pattern of haplotypes in the network (Figure 5.14) may indicate that genotyping of further markers in this clade is required to attain greater phylogenetic resolution and a better understanding of the degree of sharing of haplogroups present in Ethiopian ethnic groups. Haplogroup J1 had an MSV of 0.464, and from the network of J1 (Figure 5.16), it is apparent that there may be more derived clades amenable to SNP definition that are at high frequency, as indicated by the long branch lengths between nodes, and the occurrence of one-step neighbour clusters of haplotypes particularly in the Afar. Haplogroup A3b2b had an MSV of 0.304, however it is apparent from the long branch lengths and the one-step neighbour cluster in the Anuak in the network of haplotypes (Figure 5.12, Figure 5.13), that a more derived SNP defined clade may also be

identifiable in this ethnic group. Haplogroup E1b1a7 exhibits the lowest MSV, and consequently has the youngest estimated age of STR variation (see Supplementary Table NRYSTRdate), and intriguingly is the only haplogroup with a clear high frequency modal haplotype, and a star-like pattern (40% frequency of modal haplotype, 17% frequency of one-step neighbour haplotype to modal in E1b1a7 clade, Supplementary Table NRYaddHGSTR).

Table 5.8 Frequency of NRY haplogroups present in samples of the five ethnic groups with 15 STR haplotype information (number of samples in parentheses)

	AF	AM	AN	ML	OR	Grand Total
A3b2*(xA3b2ab)	0.00 (0)	0.00 (0)	0.03 (2)	0.02 (1)	0.00 (0)	0.01 (3)
A3b2b	0.16 (9)	0.17 (11)	0.23 (14)	0.00 (0)	0.15 (9)	0.15 (43)
B	0.00 (0)	0.00 (0)	0.02 (1)	0.00 (0)	0.00 (0)	<0.01 (1)
E*(xE1,E2)	0.00 (0)	0.02 (1)	0.00 (0)	0.00 (0)	0.00 (0)	<0.01 (1)
E1	0.54 (31)	0.40 (26)	0.02 (1)	0.88 (46)	0.66 (40)	0.49 (144)
E1b1a7	0.00 (0)	0.00 (0)	0.49 (30)	0.00 (0)	0.00 (0)	0.10 (30)
E2	0.00 (0)	0.00 (0)	0.18 (11)	0.00 (0)	0.00 (0)	0.04 (11)
J1	0.28 (16)	0.34 (22)	0.02 (1)	0.08 (4)	0.18 (11)	0.18 (54)
J2	0.00 (0)	0.05 (3)	0.02 (1)	0.02 (1)	0.00 (0)	0.02 (5)
T	0.02 (1)	0.03 (2)	0.00 (0)	0.00 (0)	0.02 (1)	0.01 (4)
No. of samples	57	65	61	52	61	296

Figure 5.11 Median joining network of 15 NRY STR haplotypes in the 296 samples with additional NRY marker information, shaded by ethnic group

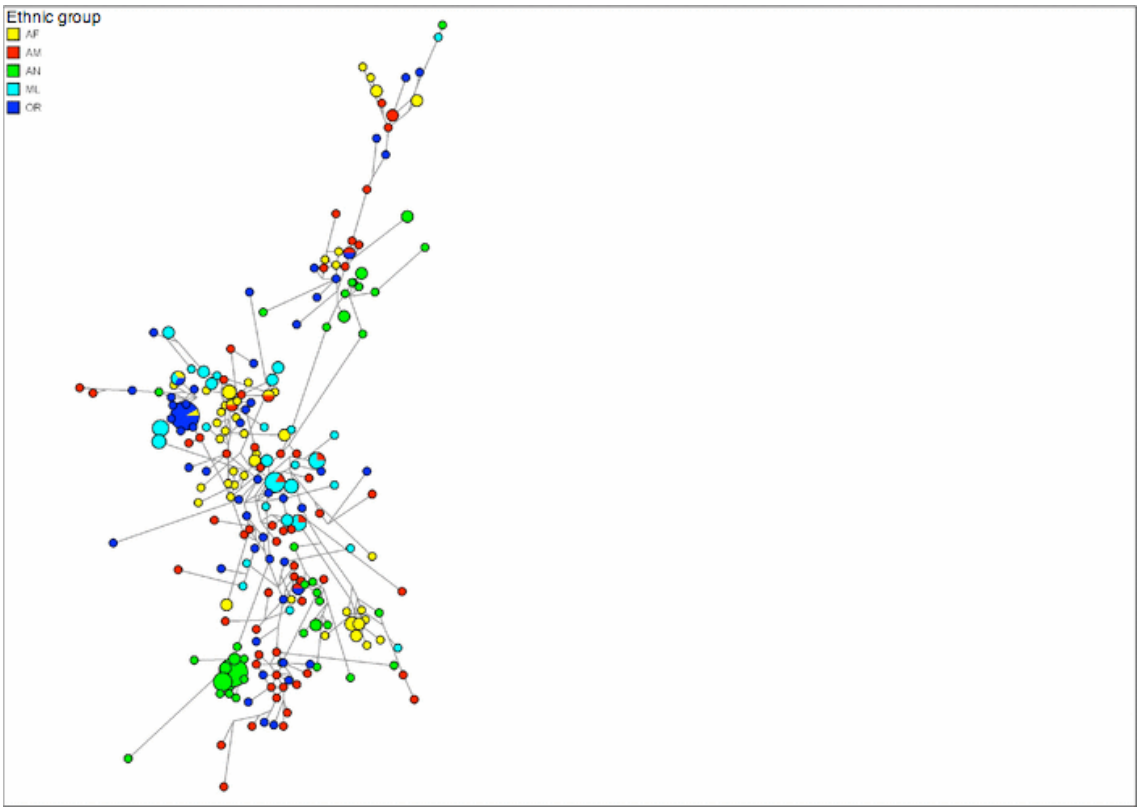


Figure 5.12 Median joining network of 15 NRY STR haplotypes in the 296 samples with additional NRY marker information, shaded by haplogroup

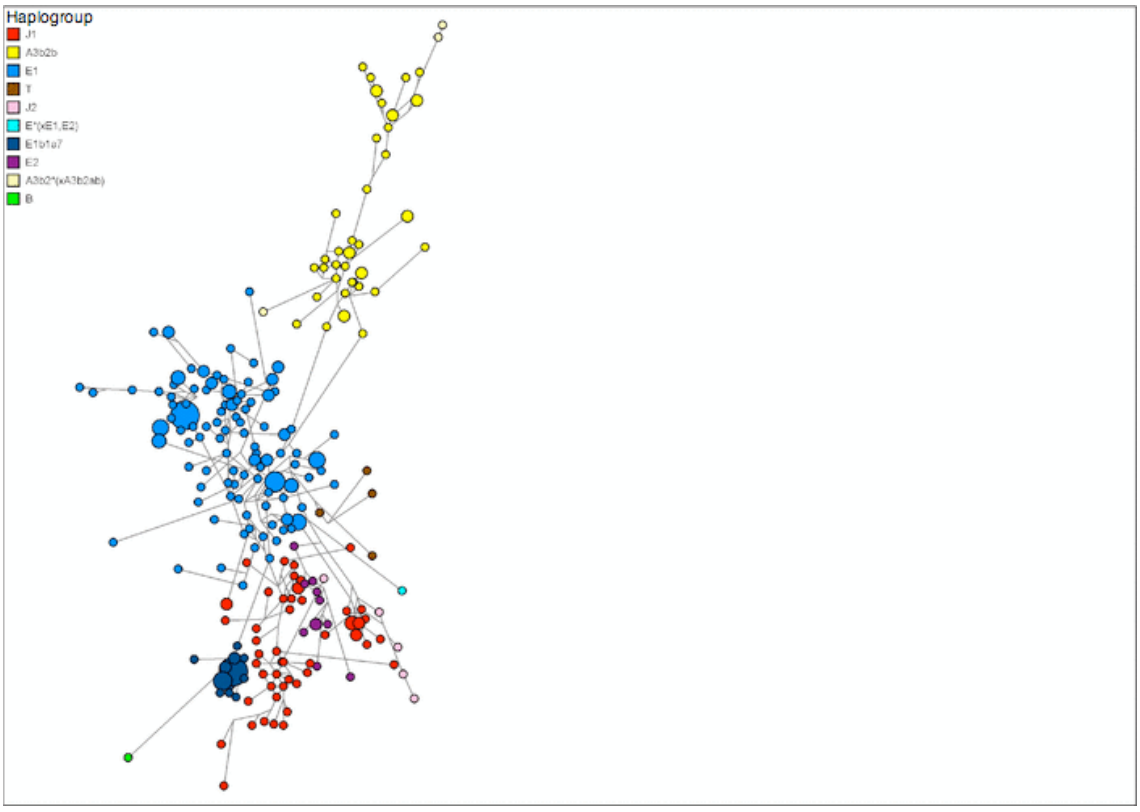


Figure 5.13 Median joining network of 15 NRY STR haplotypes in haplogroup A3b2b, shaded by ethnic group

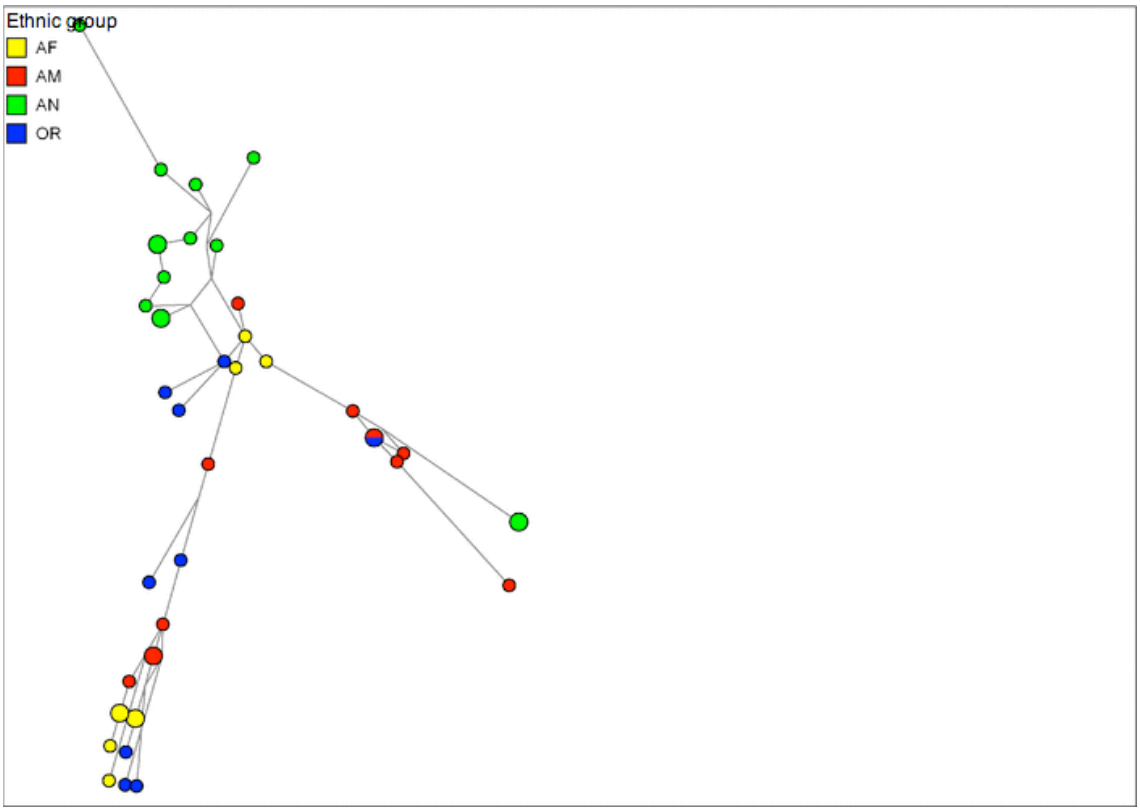


Figure 5.14 Median joining network of 15 NRY STR haplotypes in haplogroup E1, shaded by ethnic group

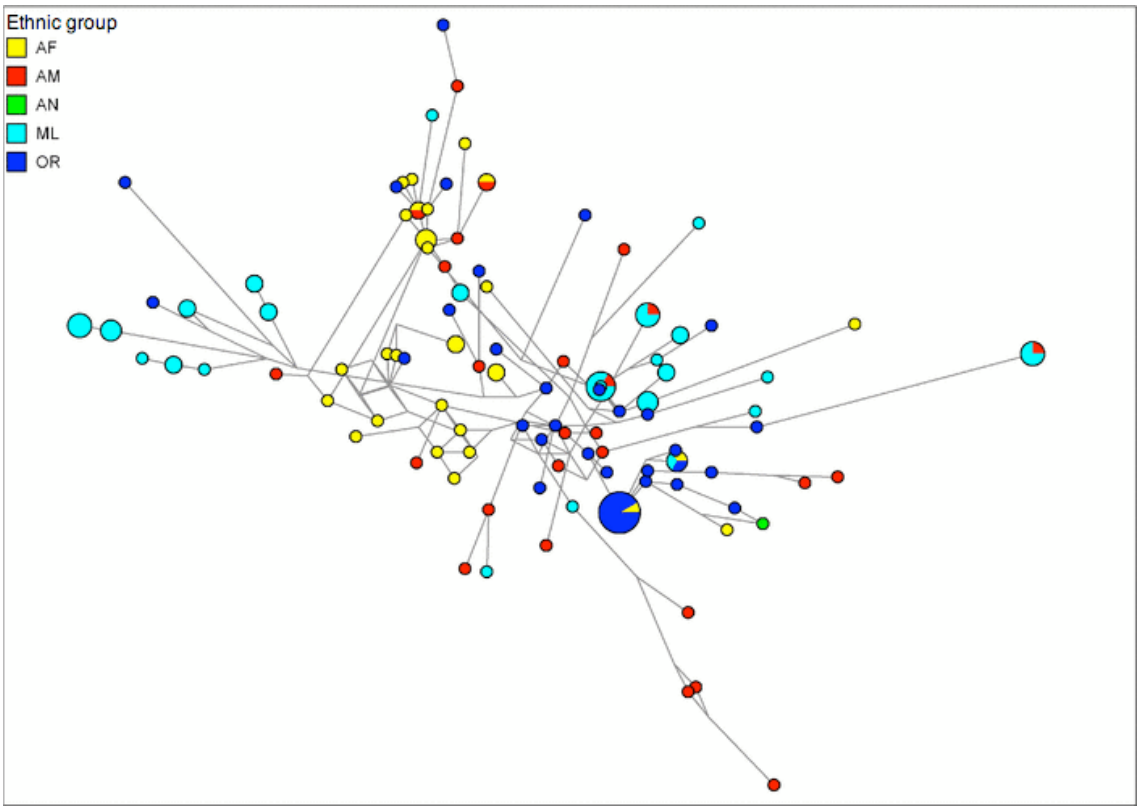


Figure 5.15 Median joining network of 15 NRY STR haplotypes in haplogroup E1b1a7, shaded by ethnic group

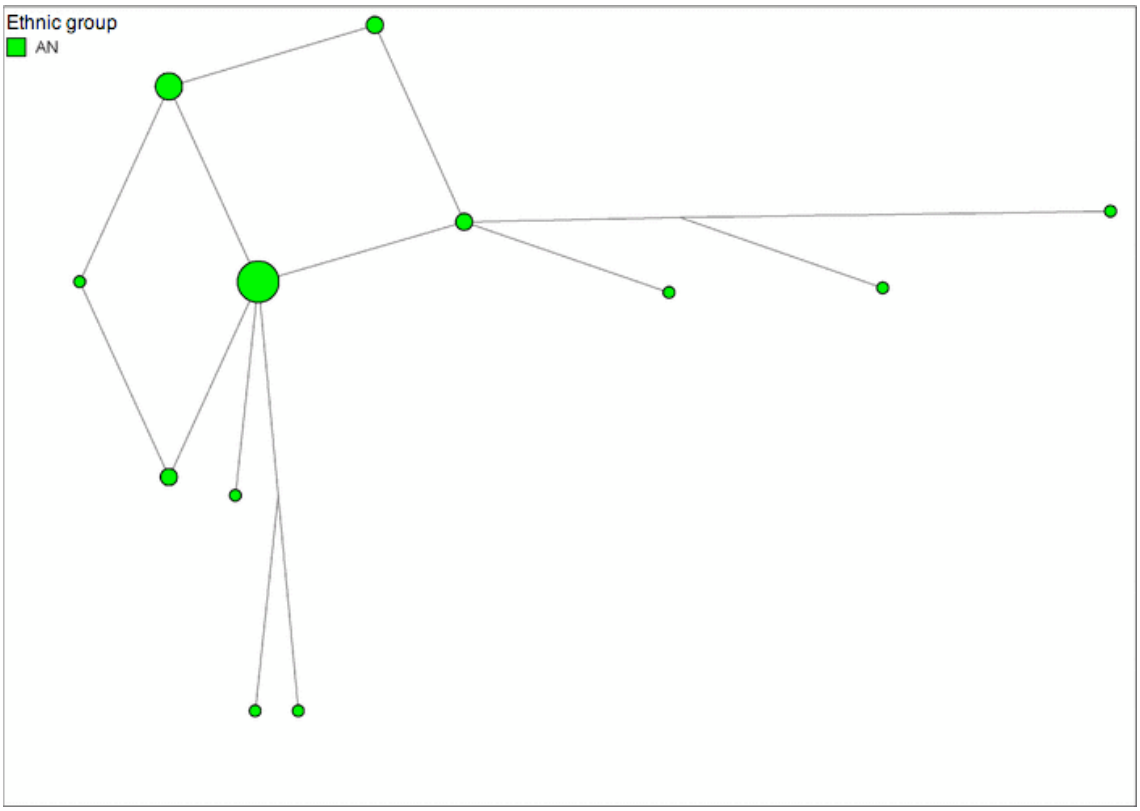
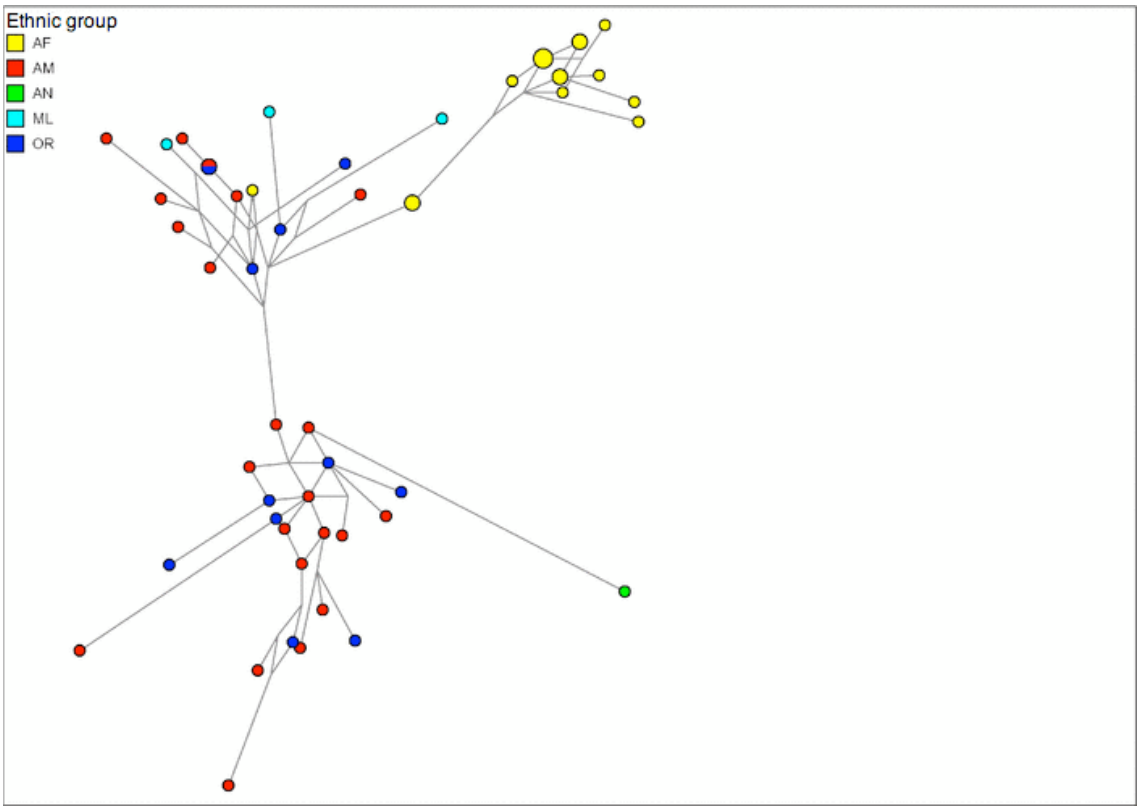


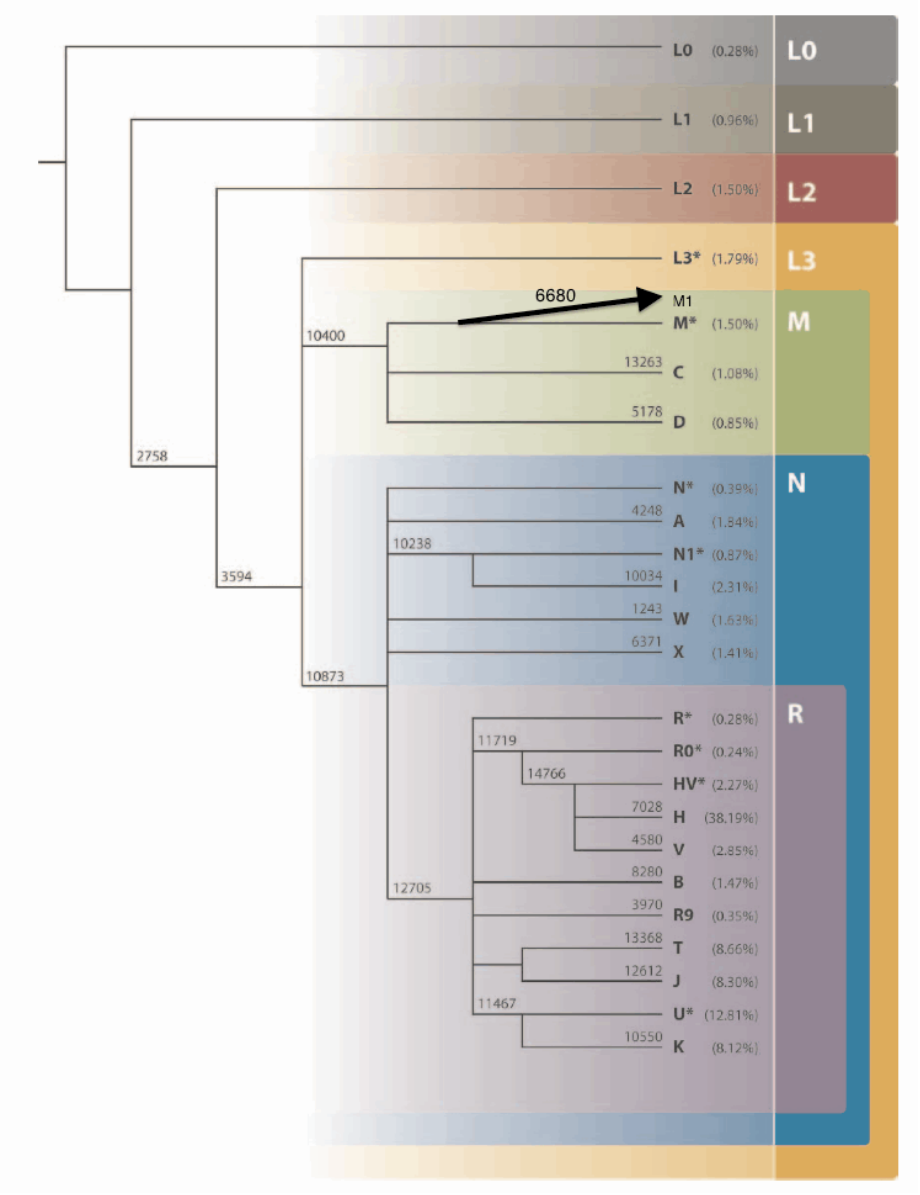
Figure 5.16 Median joining network of 15 NRY STR haplotypes in haplogroup J1, shaded by ethnic group



5.3 The distribution of mtDNA haplogroups present in five Ethiopian ethnic groups

mtDNA haplogroups were determined in 341 samples by genotyping of the 22 markers in the panel described in Behar, et al. (2007) with an additional marker for the T>C SNP at mtDNA position 6680 for the M1 clade (Figure 5.17). When combined with the HVS1 information for these samples, a single incidence of homoplasy was observed, with the same HVS1 appearing in both haplogroup H and HV* (Figure 5.19, Supplementary Table mtDNAHG).

Figure 5.17 Figure 4 from Behar, et al. (2007) showing the 22 markers and the clades they infer, with the phylogenetic location of the M1 SNP at mtDNA position 6680



The modal haplogroup was L3 in all ethnic groups except the Anuak (AN), where it was co-modal with haplogroup L2 (Table 5.9, Supplementary Table mtDNAHG). L3 occurred at highest frequency in the Maale (ML, 63%), and lowest frequency in the Amhara (AM, 25%). L2 was otherwise the second most frequent haplogroup in all ethnic groups, occurring at highest frequency in the Anuak (42%), and lowest frequency in the Maale (10%). Haplogroups L01 (a collapse of the clade L0 and L1) and M1 were the only other haplogroups observed to occur in all five Ethiopian ethnic groups. Haplogroup R0* was observed at 11% in Amhara and 4% in the Oromo (OR), but was not seen in other groups. Haplogroup K was found at 6% frequency in the Maale, and occurred at under 3% frequency in the Afar (AF), Amhara and Oromo, but was not seen in the Anuak. Haplogroups U* and N1* were the only other haplogroup found at greater than 5% frequency in an ethnic group (Afar, 6% and 7% respectively), with U* absent from both the Oromo and Anuak, and N1* absent from the Maale and Anuak. When only considering the higher level clades present in Ethiopia, the only representative of the M clade was M1, whereas 11 haplogroups belonging to the N clade were observed in Ethiopia (Figure 5.17 and Figure 5.18). Representative haplogroups of the N clade were not seen in the Anuak ethnic group, whereas N clade haplogroups were observed in the Amhara and the Afar at of 34% and 30% total frequencies respectively.

Figure 5.18 Frequency of main mtDNA haplogroups present in five ethnic groups

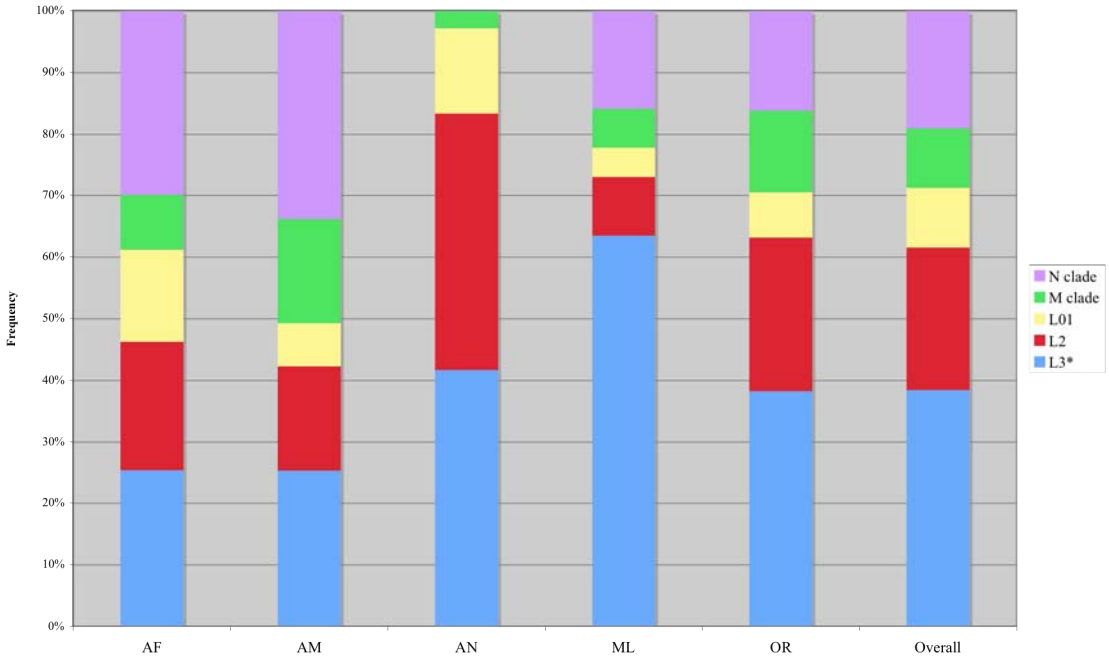


Table 5.9 Frequency of mtDNA haplogroups present in five ethnic groups (number of samples in parentheses)

Haplogroup	AF	AM	AN	ML	OR	Overall
H		0.01 (1)		0.02 (1)		0.01 (2)
HV*	0.04 (3)	0.03 (2)		0.02 (1)	0.04 (3)	0.03 (9)
J		0.03 (2)		0.02 (1)	0.01 (1)	0.01 (4)
K	0.03 (2)	0.03 (2)		0.06 (4)	0.03 (2)	0.03 (10)
L01	0.15 (10)	0.07 (5)	0.14 (10)	0.05 (3)	0.07 (5)	0.10 (33)
L2	0.21 (14)	0.17 (12)	0.42 (30)	0.10 (6)	0.25 (17)	0.23 (79)
L3*	0.25 (17)	0.25 (18)	0.42 (30)	0.63 (40)	0.38 (26)	0.38 (131)
M1	0.09 (6)	0.17 (12)	0.03 (2)	0.06 (4)	0.13 (9)	0.10 (33)
N*				0.02 (1)		<0.01 (1)
N1*	0.07 (5)	0.03 (2)			0.01 (1)	0.02 (8)
R0*		0.11 (8)			0.04 (3)	0.03 (11)
T	0.03 (2)	0.04 (3)			0.01 (1)	0.02 (6)
U*	0.06 (4)	0.04 (3)		0.03 (2)		0.03 (9)
W	0.04 (3)					0.01 (3)
X	0.01 (1)	0.01 (1)				0.01 (2)
No. of samples	67	71	72	63	68	341

The mtDNA VSOs for these 341 samples were submitted to Haplogrep (Kloss-Brandstaetter et al. 2010) (http://haplogrep.uibk.ac.at/), an automated haplogroup inference algorithm that utilises the phylogenetic data of Phylotree (version 11, http://www.phylotree.org/) (van Oven and Kayser 2009). The output from Haplogrep was then manually made comparable to the lower resolution phylogeny of Behar et al. (2007) by referring to the Phylotree (see Supplementary Table mtDNAConv). Of the 341 submitted samples, the inferred haplogroups of 317 (93%) agreed with those obtained by genotyping (Supplementary Table mtDNAInf). However, the 7% of samples whose inferred haplogroup did not agree with the genotyped haplogroup (which included 9% of all submitted VSOs), were not evenly distributed across ethnic groups or haplogroups (see Table 5.10). 99% of Oromo samples' inferred haplogroups

matched that of their genotyped haplogroup, whereas only 84% of Maale samples inferred haplogroups agreed with those obtained by genotyping. Furthermore, 12% of haplogroup L3* samples and 33% of HV* samples did not match their inferred haplogroup, and two of the rarer haplogroups, specifically, N* and X, were not correctly inferred at all. From Supplementary Table mtDNAInf, there did not appear to be a clear pattern for incorrect haplogroup assignments, and an investigation of the causes of incorrect haplogroup assignments using HVS1 data is outside the focus of this study, although it is likely that one possible cause is the lack of Ethiopian mtDNA sequences included in the Phylotree database used for haplogroup assignments.

Table 5.10 Frequency of incorrectly inferred mtDNA haplogroups present in five ethnic groups (frequency as a proportion of a haplogroup in square brackets, number of samples in parentheses)

Haplogroup	AF	AM	AN	ML	OR	Overall
HV*	0.03 [0.22] (2)			0.02 [0.11] (1)		0.01 [0.33] (3)
L3*	0.04 [0.02] (3)	0.01 [0.01] (1)	0.06 [0.03] (4)	0.11 [0.05] (7)	0.01 [0.01] (1)	0.05 [0.12] (16)
M1				0.02 [0.03] (1)		<0.01 [0.03] (1)
N*				0.02 [1.00] (1)		<0.01 [1.00] (1)
R0*		0.01 [0.09] (1)				<0.01 [0.09] (1)
X	0.01 [0.50] (1)	0.01 [0.50] (1)				0.01 [1.00] (2)
Total	0.09 (6)	0.04 (3)	0.06 (4)	0.16 (10)	0.01 (1)	0.07 (24)

Figure 5.19 Median joining network of mtDNA HVS1 haplotypes in five ethnic groups, shaded by haplogroup

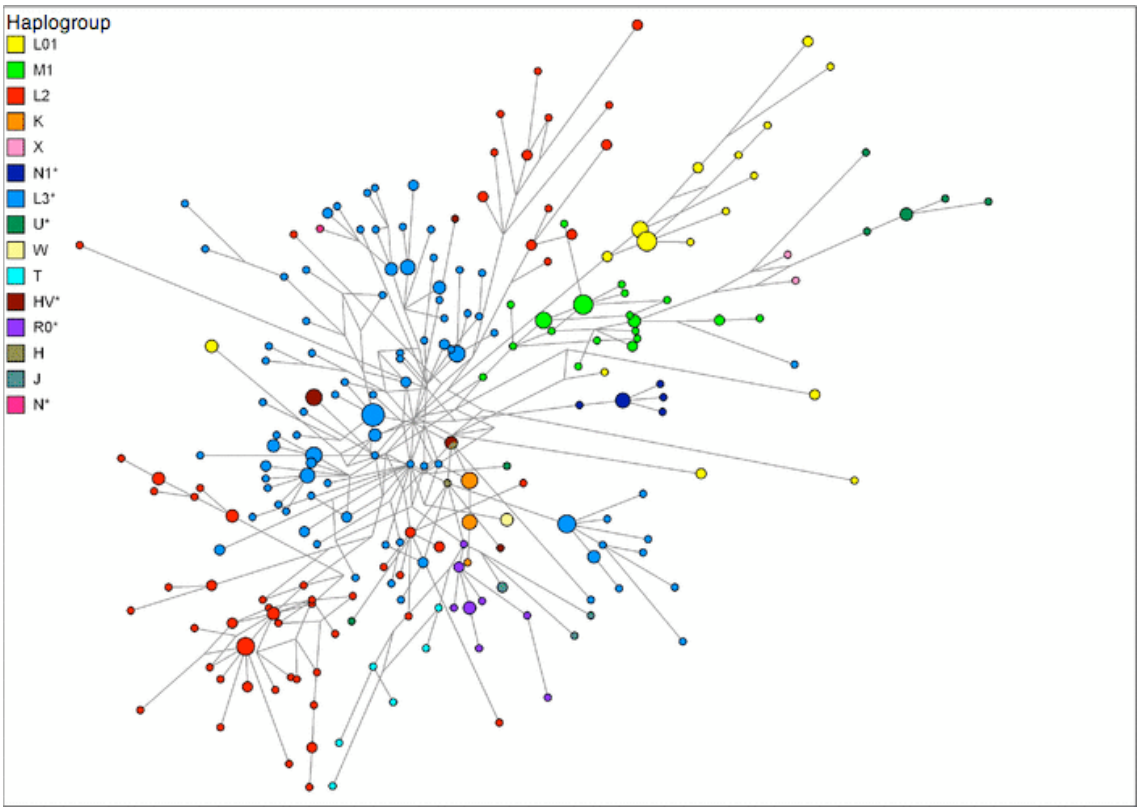


Figure 5.20 Median joining network of mtDNA HVS1 haplotypes in five ethnic groups, shaded by ethnic group

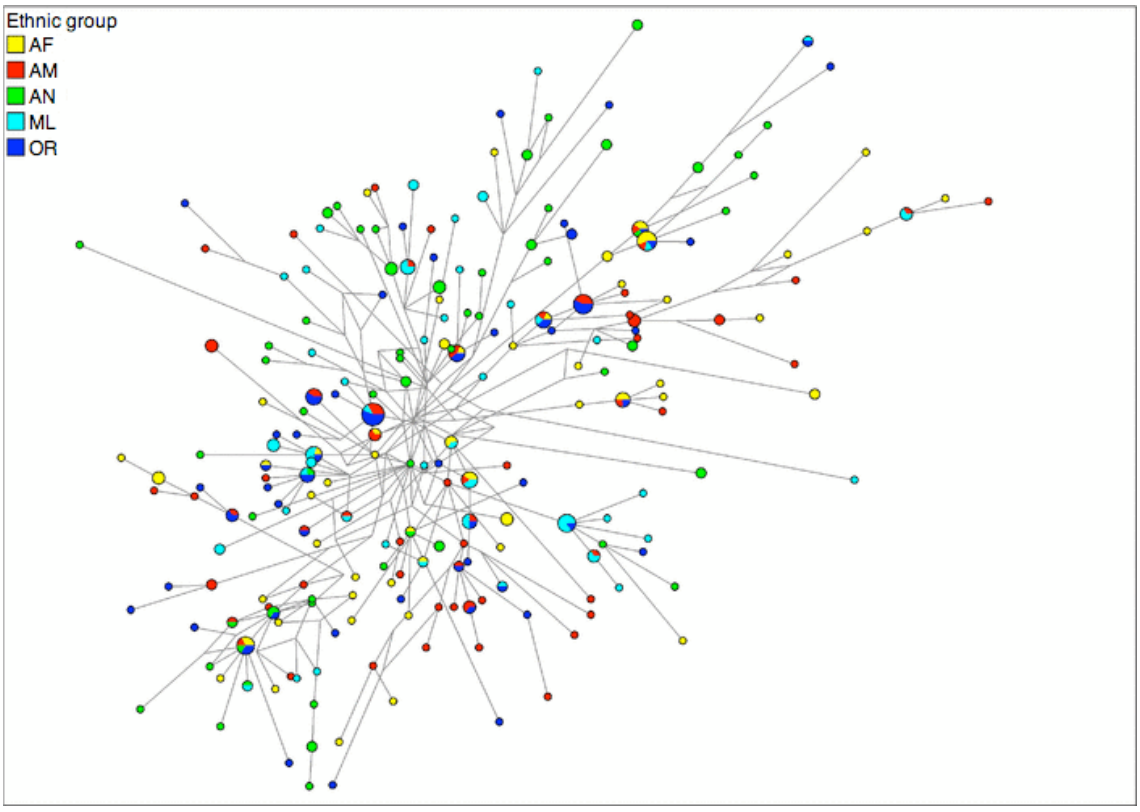


Table 5.11 shows the diversity values in the five ethnic groups. The highest haplogroup level gene diversity was observed in the Amhara (AM), with an *h* of 0.866, whereas the lowest was observed in the Maale (ML) with an *h* of 0.596. The Amhara and Maale also had the highest (0.994) and lowest (0.986) HVS1 haplotype level gene diversity respectively. The highest nucleotide diversity was observed in the Anuak (AN), with a π of 0.0297, despite having a relatively low haplogroup level gene diversity (h = 0.642), whereas the lowest nucleotide diversity was seen in the Maale (π = 0.0244).

Table 5.11 mtDNA diversity values observed in five ethnic groups

	n	Haplogroup *h*	s.d.	HVS1 haplotype *h*	s.d.	HVS1 π	s.d.
AF	67	0.858	0.043	0.991	0.012	0.0271	0.0139
AM	71	0.866	0.040	0.994	0.009	0.0251	0.0129
AN	72	0.642	0.057	0.993	0.010	0.0297	0.0151
ML	63	0.596	0.062	0.986	0.015	0.0244	0.0126
OR	68	0.774	0.051	0.991	0.012	0.0259	0.0133

In Table 5.12, it can be seen that the Anuak (AN) and Maale (ML) are distinct from all other groups while Amhara (AM), Oromo (OR) and Afar (AF) are not significantly different from each other. Comparing the ETPD results from mtDNA haplogroup frequencies (Table 5.12) to the ETPD results using the frequency of HVS1 haplotypes (Table 5.13), it can be seen that it becomes possible in the latter table to distinguish Afar from Amhara and Oromo but not the Amhara from the Oromo.

Table 5.12 P values for ETPD comparisons of five ethnic groups using frequencies of mtDNA haplogroups

	AF	AM	AN	ML
AF				
AM	0.066			
AN	<0.001	<0.001		
ML	<0.001	<0.001	<0.001	
OR	0.080	0.504	0.004	0.018

Table 5.13 P values for ETPD comparisons of five ethnic groups using frequencies of mtDNA HVS1 haplotypes

	AF	AM	AN	ML
AF				
AM	0.003			
AN	<0.001	<0.001		
ML	<0.001	<0.001	<0.001	
OR	0.004	0.690	<0.001	0.001

Figure 5.21 to Figure 5.23 show PCO plots of the genetic distances between five ethnic groups provided in Table 5.14 to Table 5.16 respectively. All three datasets show that that the greatest distance was between the Anuak (AN) and the Maale (ML), which appear at different extremities of PCO1. The Oromo (OR) consistently appear midway between these two ethnic groups, with the Amhara (AM) clustering with the Oromo when genetic distances are based on frequencies of HVS1 haplotypes (Figure 5.22, Table 5.15), indicating both the presence of haplotype sharing between Amhara and Oromo, as well as the presence of substantial non-shared haplotypes in the two groups.

Table 5.14 Pairwise Fst distances (lower diagonal) and p values (upper diagonal) between five ethnic groups based on frequencies mtDNA haplogroups

	AF	AM	AN	ML	OR
AF	*	0.276	0.005	<0.001	0.185
AM	0.003	*	0.001	<0.001	0.206
AN	0.045	0.071	*	<0.001	0.061
ML	0.094	0.095	0.111	*	0.003
OR	0.006	0.006	0.021	0.051	*

Table 5.15 Pairwise Fst distances (lower diagonal) and p values (upper diagonal) between five ethnic groups based on frequencies mtDNA HVS1 haplotypes

	AF	AM	AN	ML	OR
AF	*	0.009	<0.001	<0.001	0.004
AM	0.004	*	<0.001	<0.001	0.698
AN	0.007	0.006	*	<0.001	<0.001
ML	0.008	0.007	0.010	*	0.003
OR	0.005	0.000	0.007	0.007	*

Table 5.16 Pairwise K2P distances (lower diagonal) and p values (upper diagonal) between five ethnic groups based on frequencies mtDNA HVS1 haplotypes

0	AF	AM	AN	ML	OR
AF	*	0.015	0.002	<0.001	0.131
AM	0.015	*	<0.001	<0.001	0.194
AN	0.023	0.044	*	<0.001	0.002
ML	0.042	0.028	0.056	*	0.002
OR	0.006	0.004	0.021	0.022	*

Figure 5.21 PCO of pairwise Fst distances between five Ethiopian ethnic groups based on frequencies of mtDNA haplogroups

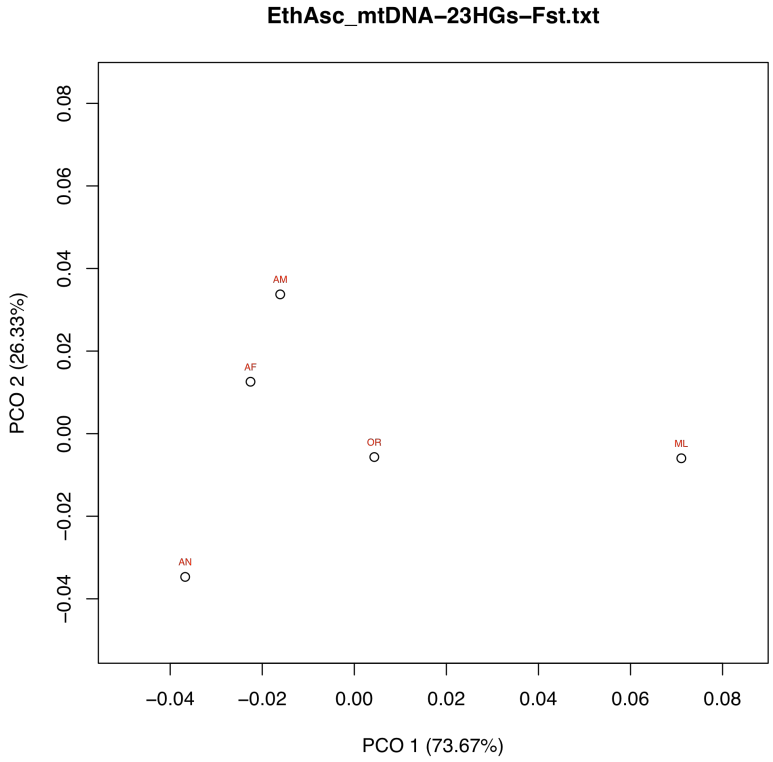


Figure 5.22 PCO of pairwise Fst distances between five Ethiopian ethnic groups based on frequencies of mtDNA HVS1 haplotypes

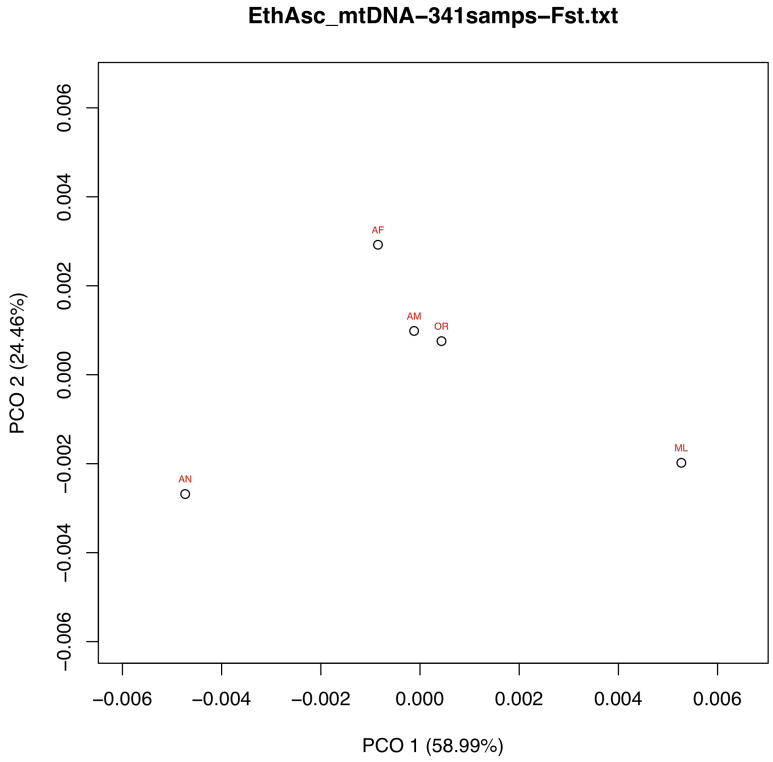
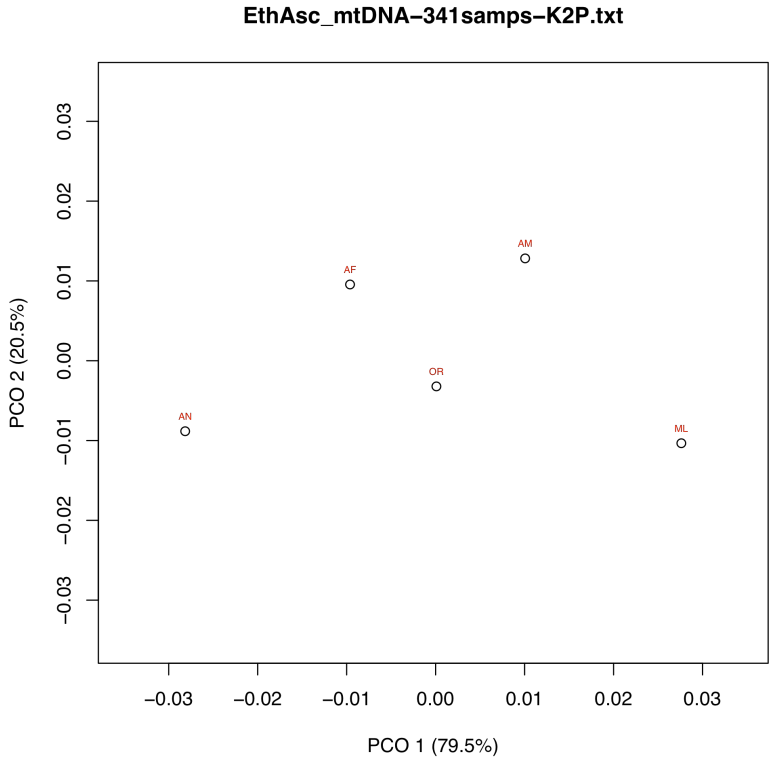


Figure 5.23 PCO of pairwise K2P distances between five Ethiopian ethnic groups based on frequencies of mtDNA HVS1 haplotypes



Chapter 6: Discussion and conclusions

6.1 Discussion and conclusions

6.1.1 The diversity and distribution of sex-specific ancestry markers in Ethiopia

The level of haplotype diversity within and amongst Ethiopian ethnic groups was highly variable for both NRY and mtDNA data. The observed levels of diversity were however comparable to that observed by Veeramah et al. (2010) in the clans of the linguistically diverse Cross River Region of Nigeria, which represents the only comparable study with respect to the number of diverse African groups surveyed in a local area for variation in NRY and mtDNA markers, and with particular attention paid to sample data collection. The Veeramah et al. (2010) study used samples collected from 24 clans, speaking six different languages (all of the Niger-Congo language family) collected at various locations in the Cross River Region. This study collected information on the sample donor's self-declared ethnic identity and clan affiliation, their current residence and birthplace, and the languages they spoke, with similar information collected for the donor's parent's and grandparent's. This enabled the authors to estimate the current level of inter-language geneflow (10%) from the proportion of sample donor's with parents that spoke two different languages. One important aim of the Veeramah et al. (2010) study was to investigate the minimum level of geneflow that had occurred between groups present in the Cross River Region during the period in which their languages diverged, consistent with the absence of significant genetic distances. The method of sample and ethnographic data collection used in this study was very similar to that used in the collection of Ethiopian samples and their ethnographic data, and the methods they used for the generation of NRY and mtDNA genetic data are also very similar to those that I have used in this thesis. In the Veeramah et al. (2010) study, even amongst linguistically diverse peoples in the Cross River Region, the vast majority of the groups had non-significant genetic distances from each other using data from NRY and mtDNA haplotypes (only 0-1.4% of NRY or mtDNA pairwise comparisons were significant at the 1% level, AMOVA NRY UEP-MS and mtDNA global Fsts both <0.001, $p>0.05$). Additionally, this study also analysed Niger-Congo speaking samples collected from three groups in the Northwest Province of Cameroon and from five groups collected in Ghana in order to put the study in the context of previous studies conducted over larger geographic areas. Significant pairwise differences at the 1% level were observed between all three datasets for both

NRY and mtDNA data, and with Ghana dataset observed to have similar homogeneity to that observed for the Cross River Region (AMOVA NRY UEP-MS and mtDNA global Fsts both <0.005, p>0.05), whereas the Cameroon dataset displayed more disparate groups (AMOVA NRY UEP-MS global Fst 0.071 and mtDNA global Fst 0.010, p<0.01 for both). The pattern observed amongst the Cross River Region groups is a stark contrast to that observed in Ethiopia (AMOVA NRY UEP-MS global Fst 0.058 and mtDNA global Fst 0.010, p<0.01 for both), with the vast majority of ethnic groups exhibiting significant genetic distances with all other groups, even at fine scales, with some genetically differentiated groups separated by less than 10km distance (for the SNNP region ethnic groups, AMOVA NRY UEP-MS global Fst 0.062 and mtDNA global Fst 0.010, p<0.01 for both). Although there is insufficient data to reach firm conclusions, it is interesting to observe that both the SNNP region of Ethiopia and the Northwest Province of Cameroon, the two areas of significant inter-group genetic distances, have, until recently, been characterised by the presence of multiple small polities, only very recently have they come under the political control of technologically more advanced societies (Freeman 2002; Freeman and Pankhurst 2003; Veeramah et al. 2008). Perhaps small, relatively isolated groups, with only moderate gene flow between them are characteristic of areas with such histories.

When Ethiopian samples were pooled according to the linguistic group of their first language (either Semitic, Cushitic, Omotic or Nilo-Saharan), all linguistic groups were observed to be significantly differentiated from each other, with Nilo-Saharan speakers substantially more differentiated from the other three linguistic groups than they were from each other, analysing both NRY and mtDNA variation. Most ethnic groups, even if they speak languages of the same linguistic group, are significantly different from each other.

When samples were grouped according the administrative region (or province) where they were collected, some similarities between groups were observed. Using the frequency of NRY UEP haplogroups (that encompass the deeper evolutionary similarities), the Oromia region was observed to be the least distinct region, with similarity with the Afar, Amhara, SNNP regions and Addis Ababa when assessed using ETPD. This is understandable in terms of the geographic area covered by the Oromia

region, which stretches across the centre of Ethiopia, and borders all of the other provinces with the exception of the Tigray administrative region. This is also consistent with the history of the Oromia peoples, and their territorial expansion and interaction with other groups from the mid 16th century AD (Pankhurst 1998; Ehret 2002). Using frequencies of NRY and mtDNA haplotypes, the two chartered cities used in this study, Addis Ababa and Dire Dawa, also showed similarities with the provinces that border them, namely the Amhara and Somali regions respectively. This is also understandable in terms of likely migration patterns, with cities often attracting diverse peoples from large areas, and representing a 'melting pot' for different ethnicities and cultures.

Despite the high degree of distinctiveness of the ethnic groups, some patterns of gene flow were discernable. ETPD revealed that there seemed to be a north-south pattern of NRY haplotype frequency similarity, connecting the Agew and the Amhara groups at its northern extent, through the Tigray, Oromo and Gurage collected in the centre of the country, and down to the Wolayta in the south. The pattern of non-differentiation using ETPD of mtDNA haplotypes revealed a similar north-south pattern of similarity, but seemed to extend further southwards and radiate out into the nearby southern groups. Ethnic groups such as Wolayta, Gamo and Dawuro demonstrated non-differentiation with many of the groups in the SNNP region, but also had links to the northern groups via the Oromo in particular. This non-differentiation amongst ethnic groups in the SNNP region could be attributable to the pre 19th Century history of the ethnic groups of these southern highlands, which was notable for conflict between several centralised polities over land and trade in the region, which would often result in vassalage of one group to another, resulting in the taking of slaves and tribute (Freeman 2002; Freeman and Pankhurst 2003). The kingdom of the Wolayta was particularly notable for their control over the trade routes through the region that ran close to Gamo territory, and would periodically expand their territory and subjugate the Gamo during periods of flourishing trade, while the Gamo would occasionally regain their sovereignty during periods of declining trade, which would then enable them to claim slaves and tribute from the neighbouring Gofa (Freeman 2002). The observed north-south pattern of similarity could be attributable to the effects of the 19th Century conquest of the south by Emperor Menilek's imperial forces and the consequent increasing influence of Amhara as a spoken language, allowing ease of movement of a) southern peoples into the northern highland areas where Amhara is more frequently spoken and b) Amhara

speaking northerners to the south where their language is increasingly becoming a *lingua franca*.

For all measures of genetic distinctiveness using NRY and mtDNA data, the Gambela region was significantly differentiated from all other areas. This highlights a general feature of the peoples of this region, namely the Anuak and Nuer ethnic groups, in that they are substantially more different from other Ethiopian ethnic groups analysed in this study. The extreme distinctiveness of the Gambela region, in an already highly heterogeneous country such as Ethiopia, could be due to a number of factors. The Gambela region has a higher frequency of Nilo-Saharan language speakers than the other provinces and is itself in a generally much more low lying area than the regions that border it (the Amhara and SNNP provinces), being separated from other Ethiopian low lying regions by the mountainous plateau to the east. Additionally, unlike many of the other groups analysed in this thesis, the Anuak and Nuer peoples inhabit both areas inside Ethiopia's borders and in neighbouring Sudan. It may be that they show far greater similarity to other ethnic groups in southern Sudan than those of Ethiopia.

The Anuak in particular, but also the Nuer ethnic group, had the highest incidence of NRY E clades, which were not observed at high frequencies in other Ethiopian groups. (Notably E1b1a7 was present in both groups and E2 in the Anuak, but not ascertained in the Nuer). Intriguingly, in a survey by Hassan et al. (2008) of 15 ethnic groups found in Sudan (that included the Nuer), haplogroups of the E1b1a clade were only observed in a single group, the Hausa, a group that is predominantly found in Nigeria (www.ethnologue.com), and haplogroup E2 was not observed at all, whereas in this thesis, haplogroup E1b1a7 was observed at 38.9% frequency in the Anuak, and 3.4% in the Nuer, while E2 was observed at 15.1% frequency in the Anuak. Haplogroups of the E1b1a clade, and in particular haplogroups E1b1a7 and E1b1a8 are most frequently found in West, Central and South East Africa, with a particular modal 6 NRY STR haplotype (15-12-21-10-11-13) considered a possible signature haplotype of the expansion of the Bantu speaking peoples (Thomas et al. 2000; Pereira et al. 2002; Veeramah et al. 2010). This STR haplotype was only observed in a single Anuak, the ethnic group with by far the highest frequency of E1b1a7. It was also observed in a single Amhara sample and two Dasanach. It is interesting that given the high frequency

of the E1b1a clade in the Anuak, and to a lesser extent, the Nuer ethnic group, that this clade was not observed in any of the indigenous Sudanese groups included in the Hassan et al. (2008) study. This could be due to the small samples sizes (all groups were 50 or less samples, 12 samples of the Nuer were included), leading to a substantial underestimate of the range of haplogroups present or that the 38.9% frequency of E1b1a7 in the Anuak is related to their particular demographic history.

NRY haplogroup A3b2 was observed at various frequencies in all of the 45 ethnic groups surveyed. A3b2 is part of the A clade, one of the deepest rooted clades in the Y chromosome phylogeny (Karafet et al. 2008). A3b2 was observed at high frequencies in the Sudanese groups in the Hassan, et al. (2008) study, with frequencies ranging up to 62% in the Dinka ethnic group. It was also observed at 33% frequency in the Nuer (consistent with the 39% frequency reported in this thesis). The highest frequency of A3b2 was 53.3%, observed in the Gedeo, a Cushitic speaking group located in the western part of the SNNP region, although other linguistically, culturally and otherwise genetically very diverse peoples also exhibited high frequencies of this haplogroup, including the Cushitc Agew (23%) of the northern highlands, the Semitic speaking Amhara and Gurage (17.2% and 22.4% respectively), the Nilo-Saharan speaking Anuak (19.4%), and the Omotic speaking Kefa and Shekecho (15.0% and 15.2% respectively). Given the widespread distribution of haplogroup A3b2 amongst the peoples of Ethiopia, this may be an indication that this haplogroup has been present at high frequencies in this region for a substantial period of time, rather than its distribution being the result of more recent introgression. Additional genotyping revealed that all of the A3b2 samples in the Afar, Amhara and Oromo, and most of the A3b2 samples in the Anuak, were of the sub-clade A3b2b, whereas the only A3b2 sample assayed in the Maale, and two of the A3b2 samples in the Anuak, did not belong to either of the known, more derived clades of A3b2, and so remain unresolved. This reflects the general lack of resolution within the A clade compared to the other clades, due to the lack of knowledge of markers that further resolve the branches within the A phylogeny (Karafet et al. 2008). Additional markers might differentiate the A3b2b clades present in the Anuak from those present in the Amhara, Oromo, and Afar ethnic groups, but for the time being this deep branching haplogroup highlights a potentially ancient relationship between otherwise very different peoples.

NRY haplogroup J was the second highest frequency haplogroup observed in Ethiopia, although unlike A3b2 and E*(E1b1a), it was not observed in every ethnic group. J clade haplogroups are frequent across North Africa, Western Asia, Europe, and the Indian sub-continent (Karafet et al. 2008 and references therein), with the sub-clade J1 more frequently observed across North Africa, the Arabian peninsular and Ethiopia, and J2 more frequently observed across Asia and Europe (Semino et al. 2004). Additional genotyping of J clades in the five ethnic groups of the Ethiopian Ascertainment panel samples revealed that the majority of J samples were of the J1 sub-clade. With the exception of the Afar (where J1 was the only observed J haplogroup), the J2 sub-clade was also observed, albeit at lower frequencies than J1, in all ethnic groups, and a single Maale sample was observed to have a J haplotype that did not belong to either of the two known derived sub-clades of J. J samples were at highest frequency (52.0%) in the Omotic speaking Shekecho ethnic group, located in the north-west SNNP region, but were also observed in other linguistically diverse and widespread groups, including the Omotic speaking Kefa and Yem (38.3% and 31.5% respectively), the Cushitic speaking Afar and Agew (25.9% and 22.3% respectively), and the Semitic speaking Amhara and Gurage (25.8% and 21.1% respectively). Notably, J clades were either not observed (Majenger) or only observed at low frequencies (1.9% in Anuak, 0.8% in Nuer) in Nilo-Saharan datasets, which is consistent with the Hassan et al. (2008) study, where haplogroup J was rarely observed in Nilo-Saharan speaking Sudanese groups. In the study by Semino et al. (2004), investigating the distribution of clades E and J, the authors interpret the distribution and internal STR variance of the J1 clade as indicative of this haplogroup spreading to Ethiopia, probably from the Western Asia, during the Neolithic period. Also, in a study by Tofanelli et al. (2009) to investigate the variation and date of the dispersal of J1, the authors ruled out the possibility that the contemporary distribution of J1 clade is due to the spread of Arab populations and Islam since the middle-ages, but concluded that it was a consequence of far more ancient events, possibly due to the movements of hunter-gatherers in response to the climactic change at the end of the Pleistocene, and mainly occurred during the mid-Holocene. Both of these scenarios would be consistent with the high level of MSV observed for J1 samples in Ethiopia, and consequently its estimated age, based on STR variation of (4,492 years using YHRD point estimate, 17,565 years point estimate using the mutation rate of Zhivotovsky et al. (2004), based on variance in 14 STRs).

NRY haplogroup E*(xE1b1a) was the highest frequency clade observed in Ethiopia, and was present at high frequencies in all groups surveyed. The high frequency of this haplogroup is, in part, a consequence of its low level of resolution, and it is likely that there would be substantial differences in frequencies of sub-clades as indicated by the large MSV for this clade (0.605 using 15 STRs). This is demonstrated by the genotyping of additional haplogroup markers. The majority of haplotypes belonged to the E1 clade in the Afar, Amhara, Maale and Oromo, whereas in the Anuak the majority of E*(xE1b1a) were found to belong to the E2 clade. Additionally, 3% of Amhara samples were observed to belong to neither the E1 nor the E2 clade, and at the present time there are no known markers that further resolve this phylogenetic position (International Society of Genetic Genealogy 2010) (Further genotyping of additional NRY haplogroup markers is currently being performed on the Ethiopian Ascertainment samples in a nested fashion to further resolve the phylogenetic position of the NRY clades in Ethiopia, and I have in this thesis, only presented the currently available data from the first round of NRY haplogroup genotyping). Notably though, the E clades present in Ethiopians are substantially different from those observed in most parts of Sub-Saharan Africa, where E1b1a clades are predominant (Jobling and Tyler-Smith 2003; Semino et al. 2004; Wood et al. 2005; Rosa et al. 2007; Sims et al. 2007; Veeramah et al. 2010).

Genotyping of mtDNA haplogroups was performed on the Ethiopian Ascertainment samples in order to determine the variation at haplogroup level in these groups. Notably, most of the major branches of the mtDNA phylogeny (L0-L3, M and N (van Oven and Kayser 2009)) were observed in Ethiopia at substantial frequencies. Haplogroups of the L series are mainly restricted to Africa, whereas the clades M and N (which are haplogroups within the L3 clade) are generally found outside sub-Saharan Africa, and are thought to only occur inside Africa due to back migration from Eurasia (Salas et al. 2002; Olivieri et al. 2006; Behar et al. 2008). Ethiopia however has previously been shown to have substantial frequencies of haplotypes of the M and N clades (Kivisild et al. 2004; Poloni et al. 2009), and the results in this thesis are consistent with previous studies. Of the M clade, only the M1 sub-clade was observed in the five Ethiopian Ascertainment groups, with highest frequency in the Amhara

(17%) and lowest in the Anuak (3%). The N clade was not observed in the Anuak, but was observed in all other groups, with the highest frequency in the Amhara (34%). Of the N clades present in the other ethnic groups, interestingly, haplogroup R0* was observed at 11% in the Amhara and 4% in the Oromo, but not observed in the Afar or Maale. The varied distribution of R0* (previously known as preHV (van Oven and Kayser 2009), which is observed at relatively high frequencies across West and Central Asia (Quintana-Murci et al. 2004)), as well as other haplogroups of the N clade, may be evidence of a more recent introgression of N haplogroups into Ethiopia (Kivisild et al. 2004). In comparison to the N clade, the distribution of the more widespread M1 haplogroup in Ethiopia (the only representative of the M clade, and present in all ethnic groups), may well be evidence of more ancient introgression from Eurasia (Gonzalez et al. 2007), which is consistent with the coalescent times estimated by Kivisild et al. (2004) for M1 (41.2 ± 17 KYA) and the (preHV)1 (14.4 ± 5.3 KYA) clades present in Ethiopians. The frequency of the mtDNA haplogroups with Eurasian origin (M and N) in ethnic groups may be associated with their proximity to West Asia and the Red Sea. The more geographically remote ethnic groups, namely the Maale and Anuak, had the lowest frequencies of M and N haplogroups compared to the more north western groups. This pattern is consistent with the Poloni et al. (2009) study analysing samples from the Nyangatom and Dasanach, peoples who are predominantly found in the remote southern lowland SNNP region, near the Kenyan border. The authors observed that the combined frequency of both clades was under 5% in each group, with the N clade not observed at all in the Dasanach. Furthermore, when these two ethnic groups were analysed alongside Tanzanian ethnic groups from the Tishkoff, et al. (2007) study, and Ethiopians from the Kivisild, et al. (2004) study, a comparison of the mtDNA genetic distances revealed that the two southern Ethiopian groups demonstrated far greater similarity with the Tanzanian groups, than they did with the northern Ethiopians. The data presented in this thesis are consistent with a general pattern of greater introgression of haplogroups of Eurasian origin into the highland ethnic groups, and also demonstrates that considerable mtDNA genetic structure exists across Ethiopia alongside the substantial heterogeneity in the pattern of NRY haplogroups.

6.1.2 The utility of additional genotyping of sex specific ancestry markers

Increasing the number of NRY STR markers from 6 to 15 greatly increased the number of private haplotypes, resulting in very little sharing of haplotypes between the five

Ethiopian Ascertainment ethnic groups. However, in comparing the genetic distances between ethnic groups determined using both the 6 STR and the 15 STR datasets, the PCO plots of both Rst and Fst distances were very similar. In both plots of Rst distances, using the two datasets, the relative positions of the five ethnic groups were almost identical and the PCOs account for very similar amounts of the overall variation, with the Anuak appearing to be the most differentiated group, positioned at the extremity of PCO1 and all the other groups placed at the opposite extremity. Despite the low level of haplotype sharing, and high level of gene diversity within the ethnic groups, the PCOs of Fst distances using the 15 STR datasets was very similar to those determined using 6 STRs, with PCO1 accounting for a similar proportion of the overall variation, and the Anuak appearing as the most genetically distant group. Using 15 STRs however, did lead to a more precise estimate for the age STR variation in NRY haplogroups, with much larger standard errors associated with the 6 STR estimate, and the ages up to 30% lower than those estimated using 15 STRs.

The use of mtDNA haplogroup data highlighted the extent of similarity between geographically and linguistically separated ethnic groups. Some haplogroups were represented in all five ethnic groups, whereas others were more restricted. In particular, the use of mtDNA haplogroup markers allows the comparison of published data on mtDNA haplogroup frequencies in Ethiopia and the rest of the world, and inferences and hypotheses to be made concerning Ethiopia's demographic history using these deeper resolution markers. However, in the use of mtDNA haplogroup frequencies to determine genetic distances and distinctions between ethnic groups, the general pattern was not dissimilar to the pattern observed using HVS1 haplotypes alone. When comparing the results of ETPD using the frequencies of mtDNA haplogroups and HVS1 haplotypes, the only difference was that the latter distinguishes the Afar from the Amhara and Oromo, whereas the former does not, with all other groups appearing to be significantly differentiated from each other. PCO plots of genetic distances using mtDNA haplogroup frequencies and HVS1 haplotypes consistently show the greatest genetic distance to be between the Anuak and the Maale, with the Oromo appearing mid-way between the two groups along PCO1, and the Afar and Amhara appearing along PCO2. Due to only a single incidence of homoplasy, with the same HVS1 sequence appearing in two different haplogroups, it should be possible to make reasonable predictions about the haplogroup status of a large proportion of the

remaining Ethiopian collection based upon their HVS1 sequence. The problem of assigning the large number of haplotype singletons to haplogroups would however remain. Haplogroup inference by reference to data from one of the most regularly updated mtDNA phylogenies (Phylotree) correctly predicted the haplogroup status of 99% of Oromo samples, but only 84% of Maale samples, with 9% of all submitted HVS1 sequences assigned to the wrong haplogroup. Consequently, due to the evolving nature of understanding of the mtDNA phylogeny, especially in Africa (Behar et al. 2008), haplogroup assignments are best made using complete mtDNA genome data. (Sequencing of complete mtDNA genomes is due to be performed on a subset of samples used in this thesis, and potentially all Ethiopian samples if the future costs of sequencing are reduced to the level where this becomes feasible). Due to the heterogeneous nature of the Ethiopian population, as demonstrated by this thesis, data from the complete sequencing of mtDNA genomes in Ethiopians could, very possibly, lead to substantial reconsideration of relationships among some of the deeper clades in the mtDNA phylogeny (Doron Behar, personal communication).

In summary, the genotyping of additional NRY STR markers and mtDNA haplogroup markers in the Ethiopian Ascertainment samples enabled far higher haplotype resolution and provided additional data to determine affinities between ethnic groups. However, the general pattern of relationships amongst these five ethnic groups using the additional NRY and mtDNA data did not appear to be substantially different to the pattern of relationships amongst the ethnic groups using the standard set of NRY UEP and MS markers, and mtDNA HVS1 haplotypes. This demonstrates the robustness of methods I have used in this thesis in determining NRY and mtDNA haplotypes, and their use in assessing the genetic affinities between Ethiopian groups.

6.1.3 Ethnic diversity in the populations studied

Diversity in the ethnic background of sample donors had substantial effects on the degree of distinctiveness of ethnic groups. If ethnically diverse samples were no different from a random set of samples in the ethnic group, then removal of these samples is unlikely to have a significant effect on the diversity of the group. It was observed however that, with the exception of mtDNA gene diversity, the proportion of samples removed was significantly correlated with the absolute change in diversity

metrics for both NRY and mtDNA data. The removal of samples with diverse ethnic backgrounds led to a significant general reduction in the NRY haplotype gene diversity in ethnic groups. However, for all other NRY and mtDNA diversity metrics, the general direction of change in diversity was not significantly different from an equal likelihood of either an increase or decrease in the metric. Consequently, in most cases, the effect of the inclusion of samples with diverse ethnic background can have significant but unpredictable effects on the overall diversity of the ethnic group.

Filtering the dataset for samples with a uniform ethnic background generally resulted in a net decrease in the number of significant pairwise ETPD comparisons, with many groups becoming more similar to each other after the removal of samples. Furthermore, genetic distances estimated using this filtered dataset demonstrated a stronger correlation with geographic distance than the original unfiltered data. Removal of these samples uncovers an underlying geographic pattern in the distribution of sex-specific lineages, and in particular when estimated using measurements that are less influenced by recent introgression (namely Fst using NRY UEP haplogroups, NRY MS Rst, and mtDNA K2P distances). This effect could be due to recent (and variable among groups) introgression into an ethnic group being primarily by individuals not local to the group, but who have an origin further afield (Amhara for example). These individuals are more likely to have sex-specific lineages that are not present in the group, or indeed any of the local groups, and therefore inclusion of these samples in the analysis tends to hide the overall deeper pattern of similarity between ethnic groups that are geographically proximate to each other. This pattern could be perceived as counter-intuitive, but it does accord with individuals mating with others outside of their ethnic group to a non-negligible degree. Due to the general relationship between geographic distance and genetics (as seen across large distances by Manica, et al. (2005) and intermediate distances by Veeramah, et al. (2010) for example), it is likely that an underlying geographically structured pattern in the genetics exists that pre-dates the ethno-genesis of the group, and that the distinctions between local groups are maintained by cultural barriers and taboos.

Comparing the removed sample set to the retained sample set using ETPD revealed that although in most cases the removed sample set was not significantly different from the

retained set, there were many instances where a significant difference was demonstrated. Instances where a substantial number of samples were removed due to diversity in ethnic background, and the removed samples set was not significantly different from the retained sample set (the Gamo for example), may indicate that introgression into a group is not necessarily a recent phenomenon, but may have been ongoing for a considerable period.

As it is not possible to predict, prior to analysis, how diversity and distinctiveness of an ethnic group may be affected by the inclusion of samples with a diverse ethnic inheritance, the ethnic identity of parents and grandparents should be considered when interpreting results of donor analysis. Donors should be chosen that are appropriate for the question under study, and detailed ethnographic information collected on each donor to understand how ethnic and linguistic diversity, as well as diversity in geographic origin, may affect the conclusions drawn from the analysis. As was seen in section 4.5, diversity and distinctiveness of ethnic groups sampled in the donor's parent's and grandparent's generations showed substantial differences, and possibly trends, across generations. If the aim of a study is to investigate the past demographic history of an ethnic group or region, and is undertaken only with respect to donor's own ethnicities, then recent introgression and changes in ethnic identity may confound the conclusions drawn. Detailed ethnographic information on ancestry should be collected from all donors as has been suggested in the past by anthropologists and linguists studying African groups (MacEachern 2000).

6.1.4 Changes in ethnicity and language over the past two generations

A wide range of values was observed in both the proportion of donors who speak the traditional language of their ethnic group and the proportion of donors for whom the ethnic group is the same as that of their parents and grandparents. Groups such as the Nuer appear ethnically and linguistically unchanged over the past two generations, whereas groups such as the Gamo show substantial amounts of change in both ethnicity and language.

Strikingly, a substantial amount of inter-ethnic marriage was observed to be taking place in Ethiopia, with on average 12.6% of sample donors with parents of two different ethnicities. Whether this level of inter-ethnic marriage is a longstanding feature of many Ethiopian ethnic groups, or whether this is a relatively contemporary phenomenon is not known, as our data does not include information on the donor's paternal grandmother and maternal grandfather. The appearance of fluidity in ethnic identity does however seem to be a more contemporary phenomenon, or at least seems to have increased in frequency over the past two generations. It was generally more likely for the ethnicity of the donor's fathers to match that of the paternal grandfather, than for the donor's ethnicity to match that of their father. Likewise, it was generally more likely for the donor's mother's ethnicity to match that of the donor's maternal grandmother than for donor's ethnicity to match that of their mother. There were many instances where the donor's ethnicity matched that of their father, but not their mother, and vice versa. However, overall there was a stronger link with the donor's ethnicity being that of their father than that of their mother. Additionally, for 2.0% of samples the donor's ethnicity had been adopted, in that it did not match either of their parent's ethnicities.

Unsurprisingly, variation in first language demonstrated similar fluidity to that observed for ethnicity, if not to a more extreme degree. 13.5% of donors had parents who spoke different first languages, and as with ethnicity, it was far more likely for the first language of the parent to be that of the respective grandparent than for the donor's first language to be the same as either parent. Notably however, there was no association of the language of the sample donor being more similar to one parent rather than the other, unlike ethnicity. The overall proportion of donors for which first language did not match that of either parent was 10.0%. This adoption of a first language that is not spoken by either parent was associated with an increase in the frequency of Amhara, and a corresponding decrease in the proportion of donors who speak the traditional language of their ethnic group. Overall, less than 84% of sample donors spoke the traditional language of their ethnic group as a first language, compared to about 96% in their parent's generation and 98% in their grandparent's generation. In four ethnic groups (the Gamo, Gofa, Gurage and Tigray) the proportion of donors who spoke their traditional first language was under 50%.

The closer linkage of ethnicity with the paternal line rather than maternal line may be due to the general practice of patrilocality of most Ethiopian marriages (Kebede et al. 2011), with greater female migration rates, resulting in greater ethnic and linguistic diversity on the donor's mother's side, matching the effects seen on genetic diversity in uniparental markers (Seielstad et al. 1998). A recent paper by Currie and Mace (2009) demonstrated that the degree of political complexity associated with a society was a strong predictor of the geographic area covered by a society's principal language. The Amhara of the northern highlands have for centuries been the dominant political and cultural group in the region, but it was only during the last quarter of the 19th century that they expanded their Empire's borders to encompass the southern provinces (Pankhurst 1998; Phillipson 1998; Zewde 2001). This territorial expansion subsequently led to increasing cultural influence of the northern highlands in the southern provinces, as well as movement of people, particularly of soldiers, settling in the south. This cultural influence was further increased by Emperor Haile Selassie during the 20th century, in his attempts to modernise the state of Ethiopia, which included the establishment of schools and universities, with the Amhara language as the primary medium of instruction, supplemented by English at higher levels, which then continued to be used during the Derg regime's mass literacy programme that ran from the 1970s until the early 1990s. After the end of the Derg regime, the Amhara language was enshrined in the 1994 constitution as the official language of the government (Hameso 1997). This spread of the Amhara language observed in the SNNP region in particular seems to be consistent with the pattern described in the Currie and Mace (2009) study.

There could be various other causes for this general pattern of change in declared ethnicity and language, not least of which is the increased mobility and means of communication in the contemporary period. Increases in mobility and quality of communication are likely to be restricted to the more developed areas, and so this pattern may actually be reflecting a greater change in ethnic identity in groups from which collections have primarily been made in towns and cities compared to collections made in remote rural areas. This is a subject that was not possible to cover in this thesis, but will be the subject of future research, as this once again highlights the importance of recording detailed ethnographic information on sample donors.

6.1.5 Associations of genetics, geography, linguistics and ethnicity

Genetic distances were observed to be highly correlated with geographic distances when metrics were used that account for the similarity between haplotypes (namely NRY MS Rst and mtDNA K2P distances), and the deeper similarity encompassed by using haplogroups. Additionally nearly all measures of linguistic and ethnic distances based upon their diversity were highly correlated with genetic distances, as well as with each other, but no correlation was observed between these distances and geographic distance. It is interesting that both the paternal ancestry specific and maternal ancestry specific genetic, linguistic and ethnic distances are generally highly correlated with each other, as this may be highlighting a major feature in Ethiopia in the general lack of sex specific processes that have shaped its diversity.

The correlation between geographic and genetic distance in this thesis is far stronger than that observed for the Veeramah et al. (2010) study, where correlations were only observed when much greater geographic separation was present. The absence of correlation between linguistic distances and geographic distances contrasts with the results of a study on the linguistically diverse island of Sumba in the Indonesian archipelago by Lansing et al. (2007), which found a strong correlation of linguistic distance with both geographic distance and genetic distance, but no correlation of genetics with geography. In this study however, linguistic distance matrices were determined by using cognates (homologous words in two or more languages that share a common ancestor), consequently inter-linguistic distances were calculated, which differs from the approach used in this thesis whereby, in the absence of similar data, distances were based upon linguistic diversity in ethnic groups. It is possible that if inter-linguistic distances were used for the Ethiopian languages in this thesis, a correlation with geography might be observed. Complete linguistic distance matrices for the languages of the 45 ethnic groups used in this thesis were not available at the time of writing, but such an analysis using this data is planned as a future project.

It is well known that the concepts regarding our understanding of language evolution are broadly similar to those that relate to biological evolution, and both can be represented using phylogenies. There have been many notable studies to date that have identified the associations of the genetic relationships amongst human populations with

the phylogenetic relationship of the languages they speak (see for example Cavalli-Sforza et al. 1988; Sokal 1988; Chen et al. 1995; Nettle and Harriss 2003; Tishkoff et al. 2009). Many of these studies have been reviewed by Diamond and Bellwood (2003), who hypothesised that these correlations reflect the movements and dispersals of the first farmers, whose languages and genes subsequently diverged together over time since their expansion from their agricultural homelands, with perhaps the most notable examples being the Bantu language speakers expanding across Sub-Saharan Africa from West Africa and the Austronesian language speakers expanding across the islands of the Pacific. The assumption that linguistic groupings and genetic groupings (including populations) evolve in parallel has however been criticised (see MacEachern 2000). Like populations, languages are not isolated units. They can be subject to substantial influences from both neighbouring and geographically distant tongues, which is often manifested as word borrowing. In addition, the language of a population can be completely replaced by that of an unrelated language, destroying any previously existing relationship between the linguistics, genetics and geography. Such language replacement could well have occurred, perhaps repeatedly, during the history of many of the contemporary Ethiopian groups included in this thesis, and this could perhaps also explain the genetic similarity amongst the linguistically very different, and geographically distant, southern Ethiopian and Tanzanian ethnic groups in the Poloni et al. (2009) study. Given the degree of linguistic change and inter-ethnic marriage in Ethiopia demonstrated by this thesis, and its effect on the sex-specific genetic systems, it is possible that if any inter-linguistic distance correlation with geographic distance once existed, then it may already have been erased.

6.1.6 Correlation between NRY and mtDNA data

Caution is advised when attempting to compare data on NRY and mtDNA variation, as these two genetic systems are not directly comparable due to their differing mutation rates and inheritance patterns, both for the systems as a whole, and variably across each genome (Underhill and Kivisild 2007). Additionally, due to the methods used in this thesis to yield information on variation in NRY and mtDNA, the numbers of different haplotypes resolved in the two systems is substantially different, with in all cases far more mtDNA HVS1 haplotypes resolved than those identified using NRY data. However, both systems are uniparentally inherited in a haploid fashion, and so both systems are sensitive to demographic changes as a consequence of their lower effective

population size compared to autosomes. Overall, it was observed that the patterns of NRY and mtDNA variation amongst the Ethiopian ethnic groups analysed in this thesis were broadly similar.

There was a remarkably strong general correlation between nearly all measures of NRY and mtDNA variation. There was a clear linear relationship between the numbers of NRY haplotypes and the numbers of mtDNA haplotypes resolved using the methods in this thesis, with approximately twice as many NRY samples per haplotype as mtDNA samples per haplotype. The same strong correlation was also observed in the haplotype gene diversity values, with a significant ($p<0.001$) ranked correlation between NRY UEP-MS *h* and mtDNA HVS1 *h*. Additionally, with the exception of NRY UEP MS Fst, Mantel correlations between NRY and mtDNA genetic distances were also highly correlated. Such a clear correlation between nearly all measures of genetic diversity suggests that both male and female lineages in most Ethiopian ethnic groups have experienced a consistent relationship to each other (although not necessarily identical demographic histories), contrasting with the sex-specific effects observed in for example the expansion of the Bantu speaking peoples (Wood et al. 2005; Pilkington et al. 2008). This pattern is shown most strongly in ethnic groups such as the Wolayta, who demonstrated both the highest level of NRY and mtDNA haplotype gene diversity, and the Majenger, who demonstrated an unusually low level of both NRY gene diversity and mtDNA gene diversity.

Such a clear general relationship amongst the 45 ethnic groups at different levels of NRY and mtDNA haplotype diversity also enables identification of ethnic groups that depart from the commonly observed pattern. The Hadiya for example exhibit an exceptionally low level of NRY haplotype gene diversity, but a moderately high level of mtDNA haplotype gene diversity. The Hadiya clearly demonstrate that the demographic histories of males and females have been markedly different from the demographic histories in other groups, with almost 50% of Hadiya samples sharing the same NRY haplotype. Similarly the Kefa exhibit one of the highest levels of NRY haplotype gene diversity but one of the lowest levels of mtDNA haplotype gene diversity.

Comparison of NRY and mtDNA genetic distances also illuminates cases where there has been a marked difference in sex specific history. One example of this is the NRY and mtDNA genetic differences between the Afar and Somali ethnic groups. Both groups practice nomadic pastoralism (Blench 1999), and both groups inhabit the eastern low-lying arid regions of Ethiopia, but with Afar occupying the northern portion, and the Somali in the south. The distribution of NRY haplotypes and haplogroups in these two groups is strikingly different, with relatively high frequencies of J haplotypes in the Afar, but very low frequencies in the Somali, and with the highest frequency of haplogroup K*(xL,N1c,O2b,P) observed in the Somali, while this haplogroup is relatively rare in the Afar. All measures of NRY haplotype genetic distances show the Somali to be one of the most genetically distinct ethnic groups, whereas the Afar were usually seen to have a degree of similarity with the northern highland groups. On the other hand, mtDNA genetic distance showed the Afar to be most similar to the Somali, (although still significantly different). A possible interpretation is that this observation reveals an underlying geographic pattern in the mtDNA haplotypes for this region that pre-dates the ethno-genesis or local establishment of these groups in contrast to the larger relative distances observed for NRY variation.

Overall, the patterns of genetic diversity and distance for Ethiopian ethnic groups using the two sex specific genetic systems were, with some exceptions, highly correlated. It is notable that far greater genetic structure was observed in both NRY and mtDNA genetic systems, over small geographic areas, than was observed in the study of the linguistically diverse people of the Cross River Region of Nigeria (Veeramah et al. 2010). In light of that study, the level of genetic diversity and distance exhibited by the ethnic groups in this thesis is remarkable given the degree of recent inter-ethnic marriage, and adoption of different languages and ethnicities in the past two generations. This pattern of diversity and underlying similarity could well have been in existence for a substantial amount of time, and will certainly be a subject of further study. It may also be that given the extraordinary level of social change that appears to have taken place over the past several generations these patterns of genetic, ethnic and linguistic distinctiveness could start to disappear.

6.2 Future work

The main aim of the research reported in this thesis was to identify the degree and distribution of diversity in sex-specific ancestry markers and ethnic and linguistic labels amongst the 45 Ethiopian ethnic groups. It is clear that far more research is required to fully appreciate the demographic histories of the peoples of Ethiopia. This thesis has barely scratched the surface in uncovering the complex interrelationships within and among the ethnic groups. As this thesis has shown, in many circumstances paternal and maternal stories show similarities, whereas in other respects they appear strikingly different. It is clear however that further research will need to be carried out sooner rather than later as the rate of change in many of these societies is increasing. The results presented in this thesis have highlighted many areas for future research, some of which are listed below:

- The 45 ethnic groups included in this thesis account for the majority of ethnic groups present in Ethiopia, and certainly account for the largest proportion of the population. Collection of samples from increasingly remote ethnic groups is ongoing however, so it will be interesting to see how these groups fit with the overall patterns of distinctiveness and similarity described in this thesis.

- It is known that there is a great degree of structure within many of the Ethiopian ethnic groups, with the presence of marginalised caste-like elements a widespread phenomenon (Freeman and Pankhurst 2003). Preliminary analysis has indicated that the culturally dominant groups and the marginalised groups can be significantly genetically differentiated, but to different degrees using the two sex-specific systems. These cases of marginalisation within ethnic groups will be investigated in the context of oral histories, and utilising simulations to explore alternative demographic scenarios.

- Due to the potential complexity of the overall demographic history of the peoples of Ethiopia, it is intended that this be investigated using simulations to see under which demographic scenarios the contemporary genetic outcomes I have described could have arisen. This will include investigations into the extent and degree of influence the peoples of the Arabian Peninsula had on the contemporary populations of Ethiopia, the genetic implications of the Oromo expansion that has taken place since the 16th century, and understanding how the

spread of both imported and indigenous agricultural practices has shaped the demography of Ethiopia.

- Applying next generation genotyping and sequencing technology to the study of genetic variation present in the peoples of Ethiopia, and see how such additional information can elucidate understanding of early human history, and the expansion of Anatomically Modern Human out of Africa.

References

(2009). Ethnologue: Languages of the World. Dallas, Texas, SIL International.

(2009). "The World Factbook 2009." from https://www.cia.gov/library/publications/the-world-factbook/index.html.

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden and I. G. Young (1981). "Sequence and organization of the human mitochondrial genome." Nature **290**(5806): 457-65.

Armitage, S., S. Jasim, A. Marks, A. Parker, V. Usik and H. Uerpmann (2011). "The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia." Science (New York, NY) **331**(6016): 453-456.

Balaresque, P., G. R. Bowden, S. M. Adams, H. Y. Leung, T. E. King, Z. H. Rosser, J. Goodwin, J. P. Moisan, C. Richard, A. Millward, A. G. Demaine, G. Barbujani, C. Previdere, I. J. Wilson, C. Tyler-Smith and M. A. Jobling (2010). "A predominantly neolithic origin for European paternal lineages." PLoS Biol **8**(1): e1000285.

Behar, D. M., S. Rosset, J. Blue-Smith, O. Balanovsky, S. Tzur, D. Comas, R. J. Mitchell, L. Quintana-Murci, C. Tyler-Smith, R. S. Wells and C. The Genographic (2007). "The Genographic Project Public Participation Mitochondrial DNA Database." PLoS Genet **3**(6): e104.

Behar, D. M., R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, S. Tzur, D. Comas, J. Bertranpetit, L. Quintana-Murci, C. Tyler-Smith, R. S. Wells and S. Rosset (2008). "The dawn of human matrilineal diversity." Am J Hum Genet **82**(5): 1130-40.

Blench, R. (1999). Why are there so many pastoral groups in eastern Africa? Pastoralists under pressure?: Fulbe societies confronting change in West Africa. A. B. Victor Azarya, Mirjam De Bruijn, Han Van Dijk. Boston, Brill Press**:** p29-49.

Cavalli-Sforza, L., A. Piazza, P. Menozzi and J. Mountain (1988). "Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data." Proceedings of the National Academy of Sciences **85**(16): 6002.

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

Chen, J., R. Sokal and M. Ruhlen (1995). "Worldwide analysis of genetic and linguistic relationships of human populations." Human biology; an international record of research **67**(4): 595.

Chiaroni, J., R. J. King, N. M. Myres, B. M. Henn, A. Ducourneau, M. J. Mitchell, G. Boetsch, I. Sheikha, A. A. Lin, M. Nik-Ahd, J. Ahmad, F. Lattanzi, R. J. Herrera, M. E. Ibrahim, A. Brody, O. Semino, T. Kivisild and P. A. Underhill (2010). "The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations." Eur J Hum Genet **18**(3): 348-53.

Cruciani, F., R. La Fratta, P. Santolamazza, D. Sellitto, R. Pascone, P. Moral, E. Watson, V. Guida, E. B. Colomb, B. Zaharova, J. Lavinha, G. Vona, R. Aman, F. Cali, N. Akar, M. Richards, A. Torroni, A. Novelletto and R. Scozzari (2004). "Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa." Am J Hum Genet **74**(5): 1014-22.

Cruciani, F., R. La Fratta, B. Trombetta, P. Santolamazza, D. Sellitto, E. B. Colomb, J. M. Dugoujon, F. Crivellaro, T. Benincasa, R. Pascone, P. Moral, E. Watson, B. Melegh, G. Barbujani, S. Fuselli, G. Vona, B. Zagradisnik, G. Assum, R. Brdicka, A. I. Kozlov, G. D. Efremov, A. Coppa, A. Novelletto and R. Scozzari (2007). "Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12." Mol Biol Evol **24**(6): 1300-11.

Currie, T. E. and R. Mace (2009). "Political complexity predicts the spread of ethnolinguistic groups." Proc Natl Acad Sci U S A **106**(18): 7339-44.

Diamond, J. and P. Bellwood (2003). "Farmers and their languages: the first expansions." Science (New York, NY) **300**(5619): 597.

Ehret, C. (2002). The Civilizations of Africa: A History to 1800. Oxford, James Currey.

Excoffier, L., G. Laval and S. Schneider (2005). "Arlequin (version 3.0): an integrated software package for population genetics data analysis." Evol Bioinform Online **1**: 47-50.

Excoffier, L., P. E. Smouse and J. M. Quattro (1992). "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data." Genetics **131**(2): 479-91.

Freeman, D. (2002). "From warrior to wife: Cultural transformation in the Gamo Highlands of Ethiopia." Journal of the Royal Anthropological Institute **8**(1): 23-44.

Freeman, D. and A. Pankhurst (2003). Peripheral People: The Excluded Minorities of Ethiopia. London, Hurst and Company.

Genealogy, I. S. o. G. (2010, 11 September 2010). "Y-DNA Haplogroup Tree 2010, Version:5.27." Retrieved 19th September 2010, from http://www.isogg.org/tree/.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman (1995). "An evaluation of genetic distances for use with microsatellite loci." Genetics **139**(1): 463-71.

Gonzalez, A. M., J. M. Larruga, K. K. Abu-Amero, Y. Shi, J. Pestano and V. M. Cabrera (2007). "Mitochondrial lineage M1 traces an early human backflow to Africa." BMC Genomics **8**: 223.

Goudet, J., M. Raymond, T. de Meeus and F. Rousset (1996). "Testing differentiation in diploid populations." Genetics **144**(4): 1933-40.

Gower, J. C. (1966). "Some distance properties of latent root and vector methods used in multivariate analysis." Biometrika **53**: 325-328.

Hameso, S. (1997). "The language of education in Africa: The key issues." Language, Culture and Curriculum **10**(1): 1-13.

Hassan, H. Y., P. A. Underhill, L. L. Cavalli-Sforza and M. E. Ibrahim (2008). "Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history." Am J Phys Anthropol **137**(3): 316-23.

Jobling, M., M. E. Hurles and C. Tyler-Smith (2004). Human Evolutionary Genetics: Origins, People and Disease. Abingdon, Garland Science.

Jobling, M. A. and C. Tyler-Smith (2003). "The human Y chromosome: an evolutionary marker comes of age." Nat Rev Genet **4**(8): 598-612.

Johanson, D. C. and T. D. White (1979). "A systematic assessment of early African hominids." Science **203**(4378): 321-30.

Karafet, T. M., F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura and M. F. Hammer (2008). "New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree." Genome Res **18**(5): 830-8.

Kayser, M., A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, L. Roewer and et al. (1997). "Evaluation of Y-chromosomal STRs: a multicenter study." Int J Legal Med **110**(3): 125-33, 141-9.

Kebede, B., M. Tarazona, A. Munro and A. Verschoor (2011). "Intra-Household Efficiency: An Experimental Study from Ethiopia." CSAE Working Paper Series.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-20.

Kitchen, A., C. Ehret, S. Assefa and C. Mulligan (2009). "Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East." Proceedings of the Royal Society B: Biological Sciences **276**(1668): 2703.

Kivisild, T., M. Reidla, E. Metspalu, A. Rosa, A. Brehm, E. Pennarun, J. Parik, T. Geberhiwot, E. Usanga and R. Villems (2004). "Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears." Am J Hum Genet **75**(5): 752-70.

Kloss-Brandstaetter, A., D. Pacher, S. Schoenherr, H. Weissensteiner, R. Binna, G. Specht and F. Kronenberg (2010). HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups.

Lansing, J. S., M. P. Cox, S. S. Downey, B. M. Gabler, B. Hallmark, T. M. Karafet, P. Norquest, J. W. Schoenfelder, H. Sudoyo, J. C. Watkins and M. F. Hammer (2007). "Coevolution of languages and genes on the island of Sumba, eastern Indonesia." Proc Natl Acad Sci U S A **104**(41): 16022-6.

Lee, A. C., A. Kamalam, S. M. Adams and M. A. Jobling (2004). "Molecular evidence for absence of Y-linkage of the Hairy Ears trait." Eur J Hum Genet **12**(12): 1077-9.

MacEachern, S. (2000). "Genes, Tribes, and African History." Current anthropology **41**(3): 357.

Manica, A., F. Prugnolle and F. Balloux (2005). "Geography is a better determinant of human genetic differentiation than ethnicity." Hum Genet **118**(3-4): 366-71.

Marshall, F. and E. Hildebrand (2002). "Cattle before crops: the beginnings of food production in Africa." Journal of World Prehistory **16**(2): 99-143.

McDougall, I., F. H. Brown and J. G. Fleagle (2005). "Stratigraphic placement and age of modern humans from Kibish, Ethiopia." Nature **433**(7027): 733-6.

Nei, M. (1987). Molecular Evolutionary Genetics, Columbia University Press.

Nettle, D. and L. Harriss (2003). "Genetic and Linguistic Affinities between Human Populations in Eurasia and West Africa." Human Biology **75**(3): 331-344.

Olivieri, A., A. Achilli, M. Pala, V. Battaglia, S. Fornarino, N. Al-Zahery, R. Scozzari, F. Cruciani, D. M. Behar, J. M. Dugoujon, C. Coudray, A. S. Santachiara-Benerecetti, O. Semino, H. J. Bandelt and A. Torroni (2006). "The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa." Science **314**(5806): 1767-70.

Panchal, M. and M. A. Beaumont (2010). "Evaluating nested clade phylogeographic analysis under models of restricted gene flow." Syst Biol **59**(4): 415-32.

Pankhurst, R. (1998). The Ethiopians: A History, Blackwell Publishing.

Passarino, G., O. Semino, L. Quintana-Murci, L. Excoffier, M. Hammer and A. S. Santachiara-Benerecetti (1998). "Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms." Am J Hum Genet **62**(2): 420-34.

Pereira, L., L. Gusmao, C. Alves, A. Amorim and M. J. Prata (2002). "Bantu and European Y-lineages in Sub-Saharan Africa." Ann Hum Genet **66**(Pt 5-6): 369-78.

Pereira, L., V. Macaulay, A. Torroni, R. Scozzari, M. J. Prata and A. Amorim (2001). "Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade." Ann Hum Genet **65**(Pt 5): 439-58.

Phillipson, D. W. (1998). Ancient Ethiopia. London, The British Museum Press.

Pilkington, M. M., J. A. Wilder, F. L. Mendez, M. P. Cox, A. Woerner, T. Angui, S. Kingan, Z. Mobasher, C. Batini, G. Destro-Bisol, H. Soodyall, B. I. Strassmann and M. F. Hammer (2008). "Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa." Mol Biol Evol **25**(3): 517-25.

Poloni, E. S., Y. Naciri, R. Bucho, R. Niba, B. Kervaire, L. Excoffier, A. Langaney and A. Sanchez-Mazas (2009). "Genetic evidence for complexity in ethnic differentiation and history in East Africa." Ann Hum Genet **73**(Pt 6): 582-600.

Population Census Commission, E. (2007). Summary and Statistical Report of the 207 population and Housing Census. Addis Ababa, United Nations Population Fund.

Quintana-Murci, L., R. Chaix, R. S. Wells, D. M. Behar, H. Sayar, R. Scozzari, C. Rengo, N. Al-Zahery, O. Semino, A. S. Santachiara-Benerecetti, A. Coppa, Q. Ayub, A. Mohyuddin, C. Tyler-Smith, S. Qasim Mehdi, A. Torroni and K. McElreavey (2004). "Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor." Am J Hum Genet **74**(5): 827-45.

Quintana-Murci, L., O. Semino, H. Bandelt, G. Passarino, K. McElreavey and A. Santachiara-Benerecetti (1999). "Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa." Nat Genet **23**(4): 437-441.

Ramachandran, S., O. Deshpande, C. Roseman, N. Rosenberg, M. Feldman and L. Cavalli-Sforza (2005). "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa." Proceedings of the National Academy of Sciences of the United States of America **102**(44): 15942.

Raymond, M. and F. Rousset (1995). "An Exact Test for Population Differentiation." Evolution **49**(6): 1280-1283.

Reynolds, J., B. S. Weir and C. C. Cockerham (1983). "Estimation of the coancestry coefficient: basis for a short-term genetic distance." Genetics **105**(3): 767-79.

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

Rosser, Z. H., T. Zerjal, M. E. Hurles, M. Adojaan, D. Alavantic, A. Amorim, W. Amos, M. Armenteros, E. Arroyo, G. Barbujani, G. Beckman, L. Beckman, J. Bertranpetit, E. Bosch, D. G. Bradley, G. Brede, G. Cooper, H. B. Corte-Real, P. de Knijff, R. Decorte, Y. E. Dubrova, O. Evgrafov, A. Gilissen, S. Glisic, M. Golge, E. W. Hill, A. Jeziorowska, L. Kalaydjieva, M. Kayser, T. Kivisild, S. A. Kravchenko, A. Krumina, V. Kucinskas, J. Lavinha, L. A. Livshits, P. Malaspina, S. Maria, K. McElreavey, T. A. Meitinger, A. V. Mikelsaar, R. J. Mitchell, K. Nafa, J. Nicholson, S. Norby, A. Pandya, J. Parik, P. C. Patsalis, L. Pereira, B. Peterlin, G. Pielberg, M. J. Prata, C. Previdere, L. Roewer, S. Rootsi, D. C. Rubinsztein, J. Saillard, F. R. Santos, G. Stefanescu, B. C. Sykes, A. Tolun, R. Villems, C. Tyler-Smith and M. A. Jobling (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.

Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay and A. Carracedo (2002). "The making of the African mtDNA landscape." Am J Hum Genet **71**(5): 1082-111.

Seielstad, M. T., E. Minch and L. L. Cavalli-Sforza (1998). "Genetic evidence for a higher female migration rate in humans." Nat Genet **20**(3): 278-80.

Semino, O., C. Magri, G. Benuzzi, A. A. Lin, N. Al-Zahery, V. Battaglia, L. Maccioni, C. Triantaphyllidis, P. Shen, P. J. Oefner, L. A. Zhivotovsky, R. King, A.

Torroni, L. L. Cavalli-Sforza, P. A. Underhill and A. S. Santachiara-Benerecetti (2004). "Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area." Am J Hum Genet **74**(5): 1023-34.

Semino, O., A. S. Santachiara-Benerecetti, F. Falaschi, L. L. Cavalli-Sforza and P. A. Underhill (2002). "Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny." Am J Hum Genet **70**(1): 265-8.

Sengupta, S., L. A. Zhivotovsky, R. King, S. Q. Mehdi, C. A. Edmonds, C. E. Chow, A. A. Lin, M. Mitra, S. K. Sil, A. Ramesh, M. V. Usha Rani, C. M. Thakur, L. L. Cavalli-Sforza, P. P. Majumder and P. A. Underhill (2006). "Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists." Am J Hum Genet **78**(2): 202-21.

Sims, L. M., D. Garvey and J. Ballantyne (2007). "Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically undefined Y-SNPs." Hum Mutat **28**(1): 97.

Slatkin, M. (1995). "A measure of population subdivision based on microsatellite allele frequencies." Genetics **139**(1): 457-62.

Sokal, R. (1988). "Genetic, geographic, and linguistic distances in Europe." Proceedings of the National Academy of Sciences **85**(5): 1722.

Sokal, R. R. and F. J. Rohlf (1994). Biometry. New York, W. H. Freeman and Co.

Stringer, C. (2003). "Human evolution: out of Ethiopia." Nature a-z index **423**(6941): 692-695.

Thomas, M. G., N. Bradman and H. M. Flinn (1999). "High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome." Hum Genet **105**(6): 577-81.

Thomas, M. G., T. Parfitt, D. A. Weiss, K. Skorecki, J. F. Wilson, M. le Roux, N. Bradman and D. B. Goldstein (2000). "Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba--the "Black Jews of Southern Africa"." Am J Hum Genet **66**(2): 674-86.

Thomas, M. G., K. Skorecki, H. Ben-Ami, T. Parfitt, N. Bradman and D. B. Goldstein (1998). "Origins of Old Testament priests." Nature **394**(6689): 138-40.

Thomas, M. G., M. E. Weale, A. L. Jones, M. Richards, A. Smith, N. Redhead, A. Torroni, R. Scozzari, F. Gratrix, A. Tarekegn, J. F. Wilson, C. Capelli, N. Bradman and D. B. Goldstein (2002). "Founding mothers of Jewish

communities: geographically separated Jewish groups were independently founded by very few female ancestors." Am J Hum Genet **70**(6): 1411-20.

Tishkoff, S. A., M. K. Gonder, B. M. Henn, H. Mortensen, A. Knight, C. Gignoux, N. Fernandopulle, G. Lema, T. B. Nyambo, U. Ramakrishnan, F. A. Reed and J. L. Mountain (2007). "History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation." Mol Biol Evol **24**(10): 2180-95.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber and S. M. Williams (2009). "The genetic structure and history of Africans and African Americans." Science **324**(5930): 1035-44.

Tofanelli, S., G. Ferri, K. Bulayeva, L. Caciagli, V. Onofri, L. Taglioli, O. Bulayev, I. Boschi, M. Alu, A. Berti, C. Rapone, G. Beduschi, D. Luiselli, A. M. Cadenas, K. D. Awadelkarim, R. Mariani-Costantini, N. E. Elwali, F. Verginelli, E. Pilli, R. J. Herrera, L. Gusmao, G. Paoli and C. Capelli (2009). "J1-M267 Y lineage marks climate-driven pre-historical human displacements." Eur J Hum Genet **17**(11): 1520-4.

Underhill, P. A. and T. Kivisild (2007). "Use of y chromosome and mitochondrial DNA population structure in tracing human migrations." Annu Rev Genet **41**: 539-64.

Underhill, P. A., G. Passarino, A. A. Lin, P. Shen, M. Mirazon Lahr, R. A. Foley, P. J. Oefner and L. L. Cavalli-Sforza (2001). "The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations." Ann Hum Genet **65**(Pt 1): 43-62.

Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino, W. H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J. R. Kidd, S. Q. Mehdi, M. T. Seielstad, R. S. Wells, A. Piazza, R. W. Davis, M. W. Feldman, L. L. Cavalli-Sforza and P. J. Oefner (2000). "Y chromosome sequence variation and the history of human populations." Nat Genet **26**(3): 358-61.

van Oven, M. and M. Kayser (2009). "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation." Hum Mutat **30**(2): E386-94.

Veeramah, K., D. Zeitlyn, V. Fanso, N. Mendell, B. Connell, M. Weale, N. Bradman and M. Thomas (2008). "Sex-Specific Genetic Data Support One of Two

Alternative Versions of the Foundation of the Ruling Dynasty of the Nso' in Cameroon." Current anthropology **49**(4): 707-714.

Veeramah, K. R., B. A. Connell, N. A. Pour, A. Powell, C. A. Plaster, D. Zeitlyn, N. R. Mendell, M. E. Weale, N. Bradman and M. G. Thomas (2010). "Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria." BMC Evol Biol **10**: 92.

Wakeley, J. (1993). "Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA." J Mol Evol **37**(6): 613-23.

White, T. D., B. Asfaw, Y. Beyene, Y. Haile-Selassie, C. O. Lovejoy, G. Suwa and G. WoldeGabriel (2009). "Ardipithecus ramidus and the paleobiology of early hominids." Science **326**(5949): 75-86.

White, T. D., B. Asfaw, D. DeGusta, H. Gilbert, G. D. Richards, G. Suwa and F. C. Howell (2003). "Pleistocene Homo sapiens from Middle Awash, Ethiopia." Nature **423**(6941): 742-7.

Willuweit, S. and L. Roewer (2007). "Y chromosome haplotype reference database (YHRD): Update." Forensic science international. Genetics **1**(2): 83-87.

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

Wood, E. T., D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, L. Louie, M. Bamshad, B. I. Strassmann, H. Soodyall and M. F. Hammer (2005). "Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes." Eur J Hum Genet **13**(7): 867-76.

YCC (2002). "A nomenclature system for the tree of human Y-chromosomal binary haplogroups." Genome Res **12**(2): 339-48.

Zewde, B. (2001). A History of Modern Ethiopia, 1855-1991, Addis Ababa University Press, James Currey, Ohio University Press.

Zhivotovsky, L., P. Underhill and M. Feldman (2006). "Difference between Evolutionarily Effective and Germ line Mutation Rate Due to Stochastically Varying Haplogroup Size." Molecular biology and evolution **23**(12): 2268-2270.

Zhivotovsky, L. A., P. A. Underhill, C. Cinnioglu, M. Kayser, B. Morar, T. Kivisild, R. Scozzari, F. Cruciani, G. Destro-Bisol, G. Spedini, G. K. Chambers, R. J. Herrera, K. K. Yong, D. Gresham, I. Tournev, M. W. Feldman and L.

Kalaydjieva (2004). "The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time." Am J Hum Genet **74**(1): 50-61.