Title Page

Human genetic variation, relationships of peoples of sub-Saharan Africa and implications for healthcare

By

Naser Ansari Pour

SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (Ph.D.)

SEPTEMBER 2010

MINOR CORRECTIONS SUBMITTED *APRIL 2011*

The Centre for Genetic Anthropology
Department of Genetics, Evolution and Environment
University College London (UCL)

Supervisor: Prof Mark G Thomas
Second Supervisor: Prof Dallas M Swallow

Declaration of Ownership

I, Naser Ansari Pour, 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.

Abstract

Sub-Saharan Africa is thought to have the most genetic variation of any continent and to be the place of origin of anatomically modern human. Nevertheless it is the subject of relatively few studies of human genetic variation. This thesis contributes to redressing this imbalance. Sex-specific genetic systems (non-recombining portion of the Y chromosome (NRY) and mitochondrial DNA (mtDNA)) along with functional nuclear loci were characterised in multiple sub-Saharan African populations with large sample sizes to infer relationships of peoples and identify implications for healthcare.

This thesis contains four projects which addressed questions in genetic anthropology, human evolution and pharmacogenetics utilising human genetic variation. In chapter 2, NRY analysis shows that a hypothesised paternal Yombe (Congo) ancestry of Palenque (Colombia), based on linguistic and historical evidence, is consistent with genetic data. Chapter 3, based on NRY data, demonstrates that a) multiple waves of migration occurred southwards during the expansion of Bantu-speaking peoples (EBSP), b) the eastern route displayed more recent migrations than the western route and c) the absence of substantial east to west NRY gene flow in sub-Saharan Africa over the past millennium. Chapter 4 suggests an eastern route out of Africa for the *CASP12* truncated variant is more likely than a western route. (The stop-codon mutation was also dated to around 120,000 YBP). Chapter 5 demonstrates that a potentially functional *CYP1A2* variant which has not been reported outside Africa is present at considerable frequencies in sub-Saharan African population groups and that exons associated with active sites in *CYP1A* genes are well conserved.

Acknowledgements

I owe a special thanks to the NERC for funding me, and to all sample donors and individuals involved with sample collections. I would also like to thank Dr Yves Moñino for collaborating with me on the Palenque project.

I also thank those individuals who have made my experience at UCL, and TCGA in particular, over the last four years an enjoyable one. This includes (in somewhat chronological order) Sarah Browning, Christopher Plaster, Krishna Veeramah, Catherine Ingram, Adam Powell, Andrew Loh, Lauren Johnson, Olivia Clark, Bryony Jones, Laura Horsfall, Rosemary Ekong, Olivia Creemer, Ripu Bains and Gurjeet Rajbans.

I am also indebted to (this time in no particular order) Professor Andres Ruiz-Linares, Professor Sue Povey and Dr Mike Weale for their invaluable guidance during the different works described in this thesis. However my main thanks are reserved for my industrial sponsor, Dr Neil Bradman, my supervisors, Professor Mark Thomas and Professor Dallas Swallow, my precious parents, Dr Mohammad-Ali Ansari Pour and Tayebeh Sedigh Orai and my adorable wife, Sajedeh Mohaghegh Hazrati. Without their support and encouragement I would have simply been unable to undertake and complete this work.

Table of Contents

List of Figures

List of Tables

Supplementary tables 2.S1 – 2.S6 are shown on the attached CD due to their large size.

Abbreviations

AA Amino-acid

AHR Aryl Hydrocarbon Receptor

AMOVA Analysis of Molecular Variance

ARMS Amplification Refractory Mutation System

ASD Average Squared Distance

cDNA complementary DNA

CDS Coding Sequence

CI Confidence Interval

CIP Calf Intestinal Phosphatase

CYP Cytochrome P450 monooxygenases

DNA Deoxyribonucleic acid

DME Drug Metabolising Enzyme

EBSP Expansion of Bantu-speaking peoples

EGP Environmental Genome Project

ELB Excoffier-Laval-Balding

EM Expectation Maximisation

ETPD Exact Test of Population Differentiation

FMO Flavin-contatining Monooxygenases

HLA Human Leukocyte Antigen

HVR-1 Hypervariable Segment 1

HWE Hardy-Weinberg Equilibrium

K2P Kimura 2 parameter

KYA Kilo Years Ago

LD Linkage Disequilibrium

MAF Minor Allele Frequency

MK McDonald-Kreitman

mtDNA mitochondrial DNA

NIEHS National Institute of Environmental Health Sciences

NRY Non-Recombining portion of the Y chromosome

PAH Poly Acyclic Hydrocarbon

PCR Polymerase Chain Reaction

PCR-M Polymerase Chain Reaction-Minisequencing

PCA Principal Component Analysis

PCO Principal Co-ordinate Analysis

RFLP Restriction Fragment Length Polymorphism

SAP Shrimp Alkaline Phosphatase

SNP Single Nucleotide Polymorphism

STR Short Tandem Repeat

TCGA The Centre for Genetic Anthropology

TMRCA Time to the Most Recent Common Ancestor

UEP Unique Event Polymorphism

UK United Kingdom

UNESCO United Nations Educational, Scientific and Cultural Organisation

US United States of America

UTR Untranslated Region

VSO Variable Site Only

YBP Years Before Present

1. Introduction

1.1. Geography, peoples and languages of sub-Saharan Africa

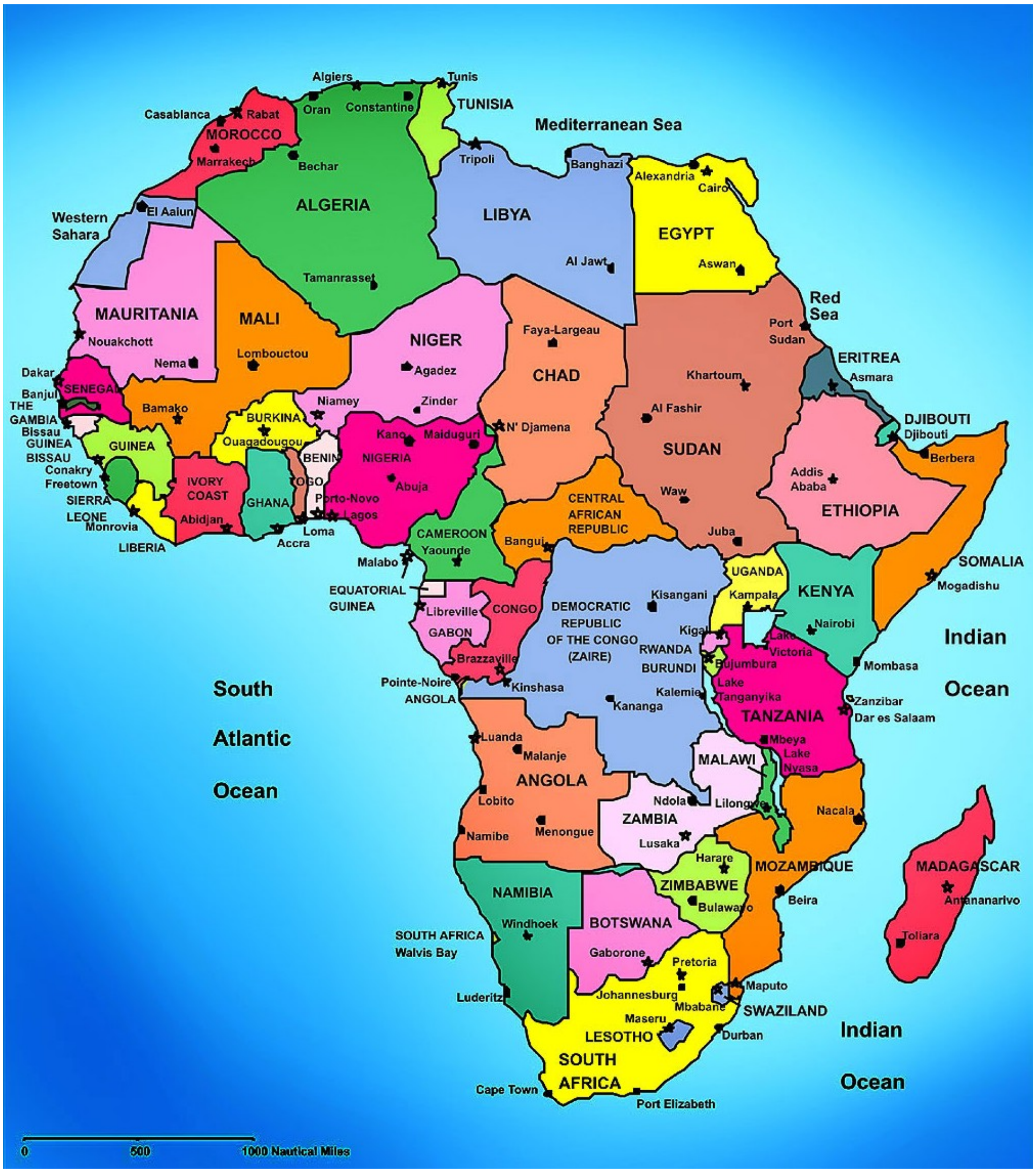
Sub-Saharan Africa is the entire African continent below the present day Saharan desert. It comprises of 45 politically defined countries (see Figure 1.1), samples from nine of which are included in this thesis. The corridor south of the Saharan desert is known as the Sahel and is a strip of semi-arid grassland stretching from the Atlantic Ocean on the west coast to Ethiopia on the east coast. Except for a few regions of highlands (e.g. Ethiopia) and deserts (e.g. Kalahari), sub-Saharan Africa is mainly comprised of tropical rainforests and wooded savannah (Cavalli-Sforza et al. 1994). Furthermore, the second largest tropical rainforest in the world is situated in the centre of sub-Saharan Africa in the Congo Basin (http://www.srl.caltech.edu/personnel/krubal/).

Africa has been a populous continent throughout the history of mankind, especially prior to the Neolithic period, with areas in which agriculture and metallurgy were pioneered showing the highest population densities in the continent. However, in the past few centuries, population growth in Africa saw a significant decline in the face of the slave trade in which the majority of slaves were shipped to the Americas (Cavalli-Sforza et al. 1994). However, in recent generations, Africa has experienced significant population growth. Currently, Africa holds approximately 1 billion individuals with over 800 million living in sub-Saharan Africa (United Nations population database; http://esa.un.org/unpp/).

Except for the 'Caucasoids' living in North Africa, the peoples of Africa have been classified into five major groups (Hiernaux 1975), although the usefulness of this classification is doubtful. The 'West African' and 'Bantu-speaking' groups, who live in the savannah and tropical forest margins, are essentially agriculturists and form the majority of the African people today. On the other hand, 'Khoisan' and 'Pygmies and

Pygmoids' groups, who live in dry and tropical areas respectively, were mostly hunter-gatherers before coming into contact with the two former groups. The relative technological development of 'West African' and 'Bantu-speaking' groups in food production has led to the heavy acculturation of hunter-gatherers and also resulted in heavy admixture between them and the agriculturists (Ehret 2002).

Figure 1.1. A political map of Africa (obtained from http://www.mapsnworld.com/)



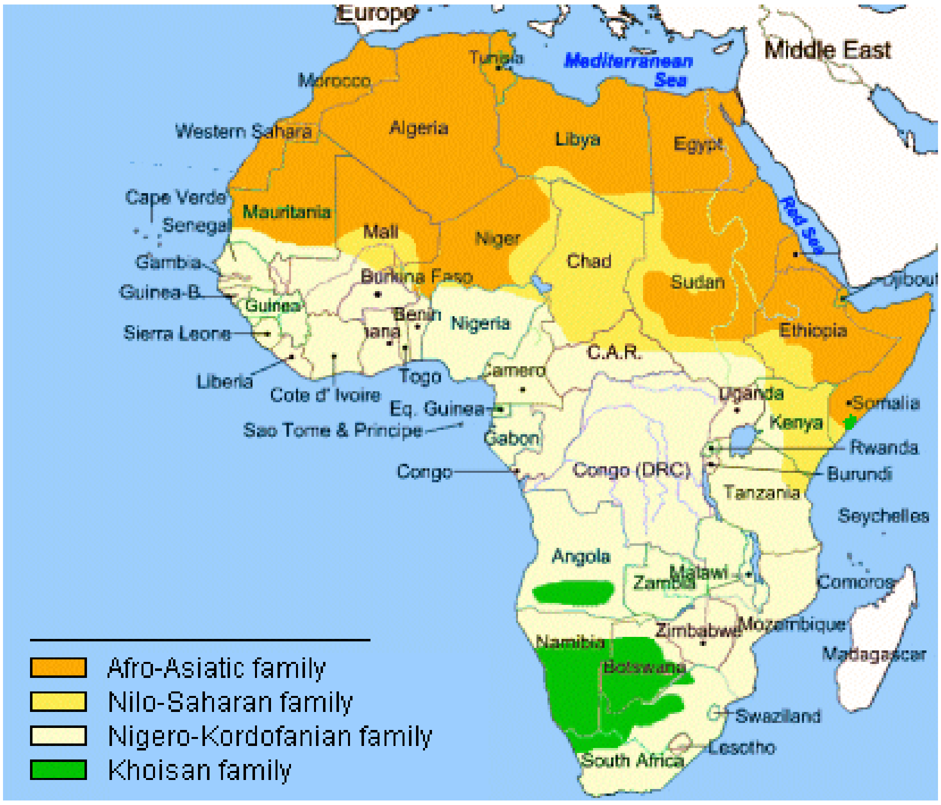
The fifth group 'Elongated Africans' are thought to be well adapted to dry and hot climates and live mostly in the savannah near the Sahel.

Africa is probably the most linguistically diverse continent in the world with nearly 1,400 languages spoken, which is approximately a quarter of all languages worldwide.
Based on the classification proposed by Greenberg (1963), all African languages can be categorised under the following four major phyla: Niger-Kordofanian, Nilo-Saharan, Afro-Asiatic and Khoisan. The Niger-Kordofanian phylum is comprised of two branches, one very large and widespread (Niger-Congo) and one small and confined to Sudan (Kordofanian) (Ruhlen 1987). It should be noted, however, according to the Ethnologue (Ethnologue 2009), Kordofanian is included in the Niger-Congo family as a sub-phylum and the classification therefore equates Niger-Congo with Niger-Kordofanian. The Niger-Congo sub-phylum is probably the largest group of languages found in sub-Saharan Africa with 1436 languages and over 350 million speakers (Ethnologue 2009). It is widely spread from Senegal in West Africa to Mozambique and South Africa (see Figure 1.2). Bantoid is a branch within the Niger-Congo family which covers about half of the languages (691 languages) and from which Bantu languages developed. The Bantu sub-group (also known as Narrow Bantu) contains 522 languages and has over 100 million speakers.

Interestingly, although Bantu languages are spoken over a wide geographic area only a small part of the diversity of the Niger-Congo family is present in this sub-group. Since it is suggested that the expansion of language families is a result of the expansion of farming populations (Bellwood 2001), the success in propagation of Bantu languages is most likely due to the expansion of its speakers. The expansion of Bantu-speaking peoples (EBSP) is thought to have happened around 5000 years ago with their advance into sub-equitorial Africa.

The Nilo-Saharan group of languages is the second largest phylum in sub-Saharan Africa with 198 languages and about 40 million speakers extending geographically from Algeria to Tanzania (Ethnologue 2009). The majority of Nilo-Saharan speaking peoples are either pastoral nomads (south Sudan) or farmers (south Chad) (Cavalli-Sforza et al. 1994).

Figure 1.2. A linguistic map of Africa according to Greenberg (1963).



Picture obtained from http://www.freelang.net/

The Afro-Asiatic family consists of 353 languages with more than 350 million speakers. Although all the languages within this phylum are not confined to North Africa (some are present in the Middle East), it only has about a fifth of the language count of the Niger-Kordafanian phylum with almost the same number of speakers. The Afro-Asiatic speaking peoples include the Berbers in Morocco, Chadic speakers in northern Nigeria, Omotic and Cushitic speakers in Ethiopia and Semitic speakers in Eritrea, Egypt, Libya and Algeria.

The smallest language phylum in the sub-Saharan region is the Khoisan family, which consists of 20 languages with nearly five hundred thousand speakers. Interestingly, almost all speakers are confined to South-West Africa (Kalahari Desert) where the EBSP did not penetrate and which has been relatively isolated for a long period of time (Denbow 1990). Khoisan languages are also known as ‘click’ languages since click sounds are used as regular consonants.

1.2. Genetic studies in sub-Saharan Africa

1.2.1. Classical genetic markers

Prior to the discovery and development of DNA genotyping techniques, genetic diversity among populations was measured based on variation observed at blood group loci (e.g. ABO blood group), Human Leukocyte Antigen (HLA) loci and blood protein electrophoretic mobility (Latter 1980; Ryman 1983). It was found that variation among major geographic regions accounts for a small percentage of total genetic variation and most of the genetic variation observed was within local populations. Cavalli-Sforza et al. (1988) looked at the variation of non-DNA markers on a much larger scale (120 alleles in 42 worldwide populations). According to the phylogenetic relationship of populations investigated, the first bifurcation in the tree separated Africans from non-Africans. Based on correlations with archaeological data, this event was dated back 100 kya, which is consistent with the time of the expansion of anatomically modern humans ‘Out of Africa’. Furthermore, detailed examination of genetic variation within the African continent was conducted by Cavalli-Sforza and colleagues and published in the book *The History and Geography of Human Genes* (1994). 49 African populations (grouped according to their country, linguistic affiliation or ethnicity) were tested for 48 genes on average. The resultant phylogenetic tree bifurcates into a large cluster representing the sub-Saharan African populations and a small cluster representing northern and eastern African populations. A second bifurcation also splits the cluster into northern and eastern sub-clusters. Within the sub-Saharan cluster two major sub-clusters are apparent where one consists of Bantu and Nilotic speaking peoples and the other represents almost all West African populations, indicating genetic differentiation between Central-Southern African and West African populations. Furthermore, based on principal coordinate analysis (PCO) results, Bantu-speaking populations, show a higher level of similarity among themselves than do West African populations indicating a more recent common origin consistent with EBSP. Except for a few deviations, the results of the analysis show that, similar to the study of worldwide populations (Cavalli-Sforza et al. 1988), a good correlation is observed between language and geography with genetics.

1.2.2. DNA markers

The discovery of polymerase chain reaction (PCR), DNA fragment analysis (RFLP), DNA dideoxy sequencing and other genotyping techniques made it possible to ascertain genetic variation directly at the DNA level. In this section, a very brief summary of autosomal, non-recombining portion of Y chromosome (NRY) and mitochondrial DNA (mtDNA) genetic markers typed in African populations is given.

1.2.2.1. Autosomal markers

The early DNA studies of African populations concentrated on the analysis of diversity, based on genome-wide RFLP and STR markers (Bowcock et al. 1994) and variation within single genes (Tishkoff et al. 1996). Until recently the pattern of genome-wide diversity across geographically and ethnically diverse African populations has been largely uncharacterized (Reed & Tishkoff 2006), which is mainly due to difficulties in collecting samples in the field. However, based on over 1300 autosomal markers typed in 2432 Africans, the genetic structure of African populations and levels of diversity have been evaluated (Tishkoff et al. 2009). Interestingly, most studies ascertaining diversity of autosomal markers conducted on African samples (even though mostly having small sample sizes), rather than functional studies, have investigated the Out of Africa hypothesis of the origins of anatomically modern humans.

Pharmacogenetic studies in African populations have been scarce compared with the number involving European and Asian (i.e. Japanese) populations (see, for example, the CYP allele database at http://www.cypalleles.ki.se/), yet have generated valuable results.

In the case of the highly polymorphic *CYP2D6* gene, Masimirembwa et al. (1993) showed stark differences in the frequency of Eurasian modal haplotypes in the Shona population (n = 76) of Zimbabwe and also noted a weak correlation between genotypes observed and phenotypic expectations (Masimirembwa et al. 1996a). The gene was further characterised by re-sequencing Shona samples and a novel SNP was found in exon 2 where a third of individuals carried this newly defined variant (*CYP2D6*17*). Interestingly, this variant explained the discrepancy observed at the phenotypic level in this African population (Masimirembwa et al. 1996b). This case study is a proof of concept that genotyping Eurasian-based diagnostic SNPs in African populations is not a safe approach in predicting phenotype variation. Therefore, it is very likely that valuable pharmacogenetic data will be generated by re-sequencing African populations of relatively large sample sizes in order to identify variants specific to African individuals that may affect phenotypic response.

1.2.2.2. Sex-specific NRY and mtDNA markers

Analysis of NRY and mtDNA markers in African populations, unlike autosomal markers, has been quite extensive. Since these two genetic systems have smaller effective population sizes compared with autosomal markers, they are more prone to genetic drift (Destro-Bisol et al. 2004; Wood et al. 2005) and, hence, more likely to vary in frequency among populations. They have been used to study genetic relationships among groups, notwithstanding that each of them represents a single, albeit complex, locus.

As shown in Cavalli-Sforza et al. (1994), mtDNA and NRY studies display genetic differentiation between North and East Africa and the remainder of Africa (see also Scozzari et al. 1999; Cruciani et al. 2002; Salas et al. 2002; Salas et al. 2004), which is in good correlation with linguistic evidence. Within sub-Saharan Africa, areas affected by the EBSP show a high level of homogeneity of both systems, with more similarity at the NRY, displaying very high frequencies of the modal EBSP NRY haplogroup (E1b1a) (Underhill et al. 2001; Wood et al. 2005; Berniell-Lee et al. 2009) (for further details on NRY and mtDNA diversity of Bantu-speaking peoples see chapter 3). Similar to areas affected by EBSP, West African populations also show a high frequency of haplogroup E1b1a (Semino et al. 2002; Wood et al. 2005; Rosa et al. 2007). This finding is of importance since this NRY haplogroup may be a signature of the EBSP (Underhill et al. 2001). However, based on the frequency of E1b1a in West Africa, it can be envisaged that this haplogroup may have been present prior to the start of the expansion and, therefore, more widely spread across sub-Saharan Africa. At the mtDNA level, however, there seems to be a higher level of differentiation between West Africa and areas affected by EBSP (Salas et al. 2002; Rosa et al. 2004). Unlike EBSP populations, East African populations (e.g. Ethiopia, Eritrea and Sudan) are genetically distinct within sub-Saharan Africa due to their high genetic heterogeneity at both NRY and mtDNA loci (Passarino et al. 1998; Underhill et al. 2001; Salas et al. 2002; Kivisild et al. 2004). One possible explanation is that, because of their geographic location, these populations have experienced substantial gene flow from North African and non-African populations (especially Arab populations). Some, at least, of the diversity may be the preservation of variation existing prior to the migration of Anatomically Modern Human 'Out of Africa'. The NRY and mtDNA of Khoisan speaking peoples have been shown to be consistently and substantially different from the same systems in all other sub-Saharan Africans (Watson et al. 1996; Knight et al. 2003). Since Khoisanids exclusively possess the most ancestral NRY and mtDNA types at high frequencies, they are thought to represent ancient remnants of paternal and maternal diversity in sub-Saharan Africa prior to EBSP and, thus, are usually the earliest major outlier branch in most phylogenetic relationships when conducting inter-population comparisons among African populations (Knight et al. 2003; Wood et al. 2005).

1.3. Rationale of the thesis:

Sub-Saharan Africa is widely accepted to be the most likely place of origin of anatomically modern humans around 200 kya (Quintana-Murci et al. 1999; Jobling et al. 2004; Campbell & Tishkoff 2008). The evidence for this comes from observations that sub-Saharan African populations have higher genetic diversity than populations outside Africa (Bowcock et al. 1994; Excoffier 2002; Tishkoff et al. 2009) and that genotypic and phenotypic (human skull) diversity within populations is negatively correlated with distance from Africa (Tishkoff et al. 1996; Prugnolle et al. 2005; Manica et al. 2007; Tishkoff et al. 2009). With nearly a third of all languages spoken and a population of about 600 million, sub-Saharan Africa has also the highest linguistic diversity on earth (Ethnologue 2009). Furthermore, due to the close relationship between language and ethnic identity, over 2000 distinct ethnolinguistic groups reside in sub-Saharan Africa (Tishkoff et al. 2009), which results in complex relationships among them.

Since very little is known about genetic variation within sub-Saharan Africa, characterisation of multiple populations will be important for understanding the history and relationships of these populations at the fine scale and identifying major demographic events at the large scale. The results will provide useful insights into whether language, geography or ethnic identity is best correlated with genetic variation and which DNA markers or genetic systems are to be tested for a specific hypothesis. Furthermore, it will also be of great significance in improving the design of pharmacogenetic studies since genes encoding drug metabolising enzymes are more likely to be harbouring more mutations or simply may have sub-Saharan-specific mutations, as in *CYP2D6*.

Most studies on sub-Saharan African populations have been based on either low sample numbers or limited number of poorly defined populations (Underhill et al. 2000; Pereira et al. 2001; Cruciani et al. 2002; Solus et al. 2004; Jiang et al. 2005; Berniell-Lee et al. 2009). In functional studies, a further methodological limitation has been the ascertainment bias towards Eurasian markers. In chapters 2-5 the issue of sample sizes has been overcome by the use of the TCGA African DNA collection, which has over 20,000 DNA samples available from many populations defined by their ethnicities across the African continent. Polymorphism ascertainment bias was also overcome by re-sequencing autosomal segments of interest in substantial numbers of African chromosomes, genotyping NRY STRs and sequencing the HVR-1 region of mtDNA. Overall, this thesis provides fine-scale data on many populations at neutral and functional loci with probably the largest and most densely assembled sub-Saharan African data set yet reported.

1.4. Overview of result chapters:

Among the four projects studied in this thesis, Chapters 2 and 3 utilise the variation observed at the NRY and mtDNA loci to address questions in genetic anthropology. Chapter 4, similar to the approach taken by Tishkoff et al. (1996), focuses on the variation within *CASP12* to infer the most likely route of the truncated allele spreading out of Africa. Chapter 5 investigates genetic variability at both *CYP1A* loci through re-sequencing and SNP genotyping and identifies implications for healthcare in sub-Saharan Africa.

Chapter 2 uses sex-specific genetic data, which are used to examine the origins of a Colombian rural community (Palenque) who are believed to be descendants of a group of slaves originating from Africa, specifically Yombe, from Congo, a few centuries ago. In this investigation, it was shown that the distribution of the paternally inherited NRY variation in the Palenque was most similar to Yombe compared with the other 41 sub-Saharan African populations, all of which could have contributed individuals to the Atlantic slave trade (2870 samples; median sample size of 54), including representatives of seven other Congolese groups. The NRY data supported the linguistic evidence that the Yombe were the most likely founding group of Palenque.

Chapter 3 utilises the same set of sex-specific genetic data generated for 42 sub-Saharan African populations in chapter 2 (plus a population in Ethiopia (Anuak)) to shed further light on the expansion routes taken by the EBSP. To the best knowledge of the author, hitherto, genetic studies of the EBSP have been generating low resolution data even though the EBSP modal haplogroup has been further characterised (Karafet et al. 2008). Based on 2800 sub-Saharan African samples (43 populations), the E1b1a haplogroup was further characterised and new insights regarding the routes of the EBSP have been obtained.

Chapter 4 is based on *CASP12* variation in sub-Saharan Africa and investigates the most likely route of the truncated allele (carrying the stop-codon in exon 4) spreading out of Africa (Xue et al. 2006). Based on 534 samples (five African, four European, two Middle

Eastern and one East Asian populations), sub-Saharan African populations had fewer truncated haplotypes and exhibited higher haplotype diversity than North African and non-African populations. Patterns of haplotype diversity suggest that the truncated variant most likely left Africa via an eastern route into Eurasia. This is in good agreement with other studies suggesting a major eastern route of dispersal of Anatomically Modern Human 'Out of Africa'.

Chapter 5 examines the distribution of genetic variation in Africa of two genes that encode drug metabolising enzymes Cytochrome P450 1A1 and 1A2 (*CYP1A1* & *CYP1A2*). It describes the variation observed in both genes and reports levels of conservation at both loci. Similar to previous studies (Masimirembwa et al. 1996b; Veeramah et al. 2008b), many of the variants characterised in this study were African-specific. Analysis of diversity in the two *CYP1A* genes in humans and their close primate relatives suggests that while there are strong forces acting to conserve the gene sequences, the diversity displayed by the peoples of sub-Saharan Africa, and Ethiopia in particular, presents an opportunity to research the detailed metabolic functions of the enzymes, which is not well understood.

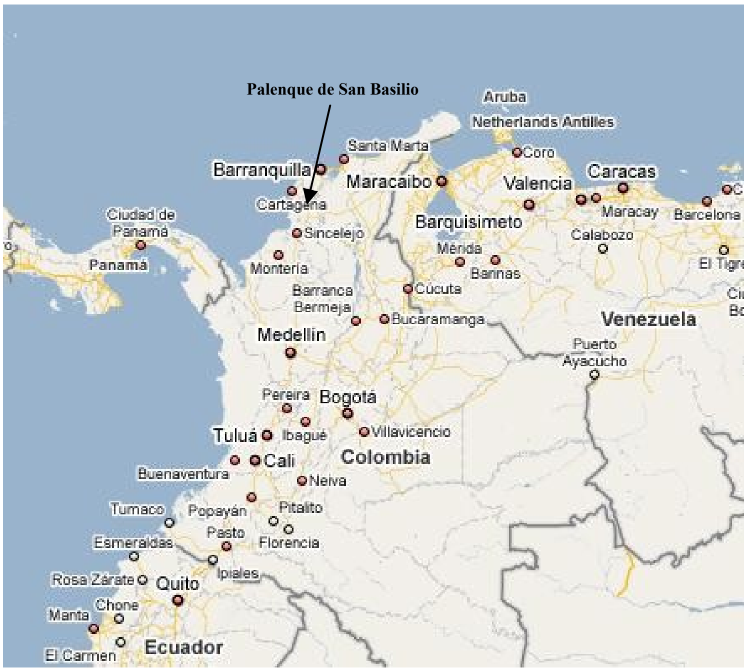
2. The origins of the Palenque: a Kikongo speaking community in rural Colombia

2.1. Introduction

2.1.1. The history and language of Palenque

During the Atlantic slave trade, in many locations throughout the Caribbean and Latin America, runaway slaves established fortified villages in their pursuit of freedom. In Colombia, these walled towns (known as palenques) were famed for their resistance to the Spanish military. This is evident from colonial records which tell the story of inhabitants successfully repulsing attacks by the authorities (Wade 1995). Despite their resistance, of the many palenques that existed in Colombia during the past few centuries, Palenque de San Basilio (see Figure 2.1) is claimed by UNESCO to be the only one that survives and in 2005 they designated the village a "Masterpiece of the Oral and Intangible Heritage of Humanity" (www.unesco.org/culture/ich).

Figure 2.1. Geographic location of the village of Palenque de San Basilio in Colombia.



Picture obtained from Google map (www.maps.google.com).

The people of Palenque de san Basilio (Palenque for short) are a community of about 3,500 individuals living in two major districts, Arriba and Abajo. Their village is located some 70 km south east of the regional capital of Bolivar, Cartagena, in north-west Colombia (10.1°N, 75.2°W). They have remained largely isolated from the prevailing Hispanic culture, living by subsistence farming together with a little cattle raising (Bickerton & Escalante 1970). Their oral history is of descent from a group of male slaves who escaped captivity from nearby Cartagena (then Latin America's major slave trade centre (Del Castillo 1984)) early in the 17th century.

Another interesting feature of the Palenque is that they are the only Colombian black community that speaks a Creole Spanish known as Palenquero (Schwegler 2006). Linguists originally held the opinion that the early Palenque settlement was formed from an amalgamation of peoples of different African ethnicities and, therefore, spoke a mixture of languages (Schwegler 2009). This opinion was based mainly on the belief that in the 17th century, Cartagena held a large number of slaves with different ethnolinguistic backgrounds (Heywood & Thornton 2007). However, subsequent analysis of linguistic data led to the suggestion that the language of the founding group originated in the area of present day Congo and/or northern Angola (Bickerton & Escalante 1970; Granda 1971).

More recently, thorough lexical research has established that Palenquero contains more than 200 words of African origin (Schwegler 2000 and 2002). Although the initial set of data had a few discrepancies, further research carried out over the past three decades claims that Kikongo is the only demonstrable donor of the African vocabulary in the Palenquero creole (Schwegler 2009). It is claimed that Palenquero, since its birth, has been locally preserved because of its relative geographic and cultural isolation (Schwegler 2006). Another factor that may have reinforced homogeneity of ethnic identity in Palenque is an 18th-century law in Colombia which granted the residents of Palenque self-government and freedom providing that they did not accept fugitives or white people into their community (Bickerton & Escalante, 1970).

In light of new linguistic data on the Kikongo origins of Palenquero, re-evaluation of the historical data was appropriate. Analysis of historical records indicated that, unlike the oral tradition, the establishment of Palenque had most likely occurred in the second half of the 17th century since slaves from the Kongo region in West Africa first started to arrive in Cartagena in large numbers in 1640. Consequently, it was no earlier than 1650 that Kikongo speakers had a significant population size to create a group of rebels who would attempt to escape captivity (Del Castillo 1984).

The Kikongo group of languages encompasses several tongues, including Bembe, Kuni, Lari, Sundi, Vili and Yombe, which are all still extant, being spoken by many people in the Republic of Congo (Ethnologue 2009). Linguists have attempted to narrow down the most likely African root of Palenquero among the Kikongo languages. Although the recorded vocabulary in Palenquero does not suggest preference for any of the languages, the ritual vocabulary (Schwegler 2009) and historical evidence (Monino 2007) suggest that Yombe is possibly the source. Today, the Yombe tongue is spoken by the Yombe people, an ethnic group mainly living in Pointe-Noire (Republic of the Congo).

2.1.2. Genetics and origin of African Diasporas

Understanding the origins of people has always been fascinating and genetics has been shown to be very useful in revealing origins through DNA analysis (Jobling et al 2004). Variation at the non-recombining portion of the Y chromosome (NRY) and mitochondrial DNA (mtDNA) has been used in genealogical inquiries of paternal (Foster et al. 1998) and maternal (Ivanov et al. 1996) lineages respectively, with conclusive outcomes. This type of analysis has the potential for use in tracing the descendants of historical Diasporas and has proved quite convincing, especially in the case of Jewish populations (Thomas et al. 1998, Thomas et al 2000 and Thomas et al. 2002).

Geographic origins of African Diasporas, in particular those created by the Atlantic slave trade, have also been investigated using sex-specific genetic systems (NRY and mtDNA) along with functional autosomal markers. NRY and mtDNA variation of peoples of Cape Verde Islands was used to detect their paternal and maternal origins respectively (Goncalves et al. 2003 and Brehm et al. 2002). Although data obtained from Islands of Cape Verde indicate a homogeneous African maternal contribution, NRY shows high variability of lineages, with less than a fifth of individuals carrying the expansion of Bantu-speaking peoples (EBSP) modal haplogroup (E1b1a). This is an interesting finding because even though Cape Verde is in close proximity to West Africa, E1b1a is not the predominant haplogroup and the NRY pool reflects a greater European, rather than West African, descent.

Genetic characterisation of peoples of San Tomé Island (Gulf of Guinea) is another example of investigation into origins of African Diasporas. Autosomal markers (Tomas 2002), mtDNA (Trovoada et al. 2004) and NRY (Goncalves et al. 2007) have been analysed to assess the extent of the African genetic contribution to the San Tomé Island population. Autosomal markers show approximately 90% genetic input from West/Central Africa, which is in agreement with the historical evidence of the major geographic sources of slaves. Furthermore, NRY data show that 84% of individuals carry the Bantu modal haplogroup (E1b1a) and variation in mtDNA indicates no European contribution. Overall the data is conclusive that, unlike Cape Verde Islands, the majority of people in San Tomé Island are of African descent and early enslaved settlers had an approximately equal sex ratio.

Genetic analysis has also been used to research the origins of Palenque (Jimenez et al. 1996; Arnaiz-Villena et al. 2009). Both studies typed HLA autosomal markers and found that the distribution of HLA antigens in Palenque is closer to that seen in West Africans than to other Colombians. Although the HLA data suggest an African ancestry component for Palenque, no study has previously been undertaken analysing NRY and mtDNA.

In all cases mentioned here, each group can be traced back to Africa and sometimes to broad geographical regions within Africa in concordance with historical information.

However, no attempt has been made to trace ancestry to specific localised regions or identified ethnic groups within Africa. This is primarily due to the lack of high-resolution genetic data in these studies, while the high level of genetic homogeneity, low coverage and patchy sampling in Africa are also key factors. As a consequence, at present, adopting genetic markers as signatures of a particular geographic location or ethnic identity in Africa should be undertaken with considerable caution.

2.1.3. Expectations of sex-specific genetic variation in the Palenque and EBSP

The Palenque are thought to have been in cultural and geographic isolation for most of their existence. However, during the past few decades, they have experienced more contact with people from outside their group (Schwegler 2009). Therefore, in recent times, an increased level of introgression is plausible. Assuming that the founding group were of recent African descent, genetic analysis can establish whether introgression from other groups has subsequently occurred or whether the Palenque have been in isolation from people living in the vicinity of the village (especially the Spanish). In the presence of little gene flow between the Palenque and the people outside, it can be expected that variation in NRY is low and that most chromosome types will be of those found in sub-Saharan Africa, with a minor fraction of European and Hispanic types. As the oral tradition only covers the paternal ancestry of Palenque, no prior hypothesis involving their maternal ancestry, beyond an assumed sub-Saharan ancestry, was constructed. From colonial records, it is evident that in the second half of the 18th century, 178 black families occupied Palenque (Del Castillo 1984). Therefore, it can be envisaged that the majority of the females at that time had a recent African origin.

Sub-Saharan Africa is known for its high genetic diversity (Bowcock et al. 1994; Kaessman et al. 1999; Yu et al. 2002). However, this is not the case for NRY variation in areas affected by EBSP (Hammer et al. 2001). It is thought that the expansion, which was the result of elevated population growth following the introduction of farming, has brought about considerable similarity among sub-Saharan African population groups (Cruciani et al. 2002; Pereira et al. 2002). Genetic data, to date, indicate that population groups in sub-Saharan Africa have a high proportion of E1b1a (previously known as E3a and hg8) samples (Pereira et al. 2002; Beleza et al. 2005; Rosa et al. 2007) and also have a characteristic common modal NRY STR haplotype (Thomas et al. 2000; Pereira et al. 2002), indicating a high level of patrilineal genetic homogeneity. Furthermore, studies of mtDNA variation among sub-Saharan African population groups have also shown a pattern of considerable genetic similarity (Pereira et al. 2001; Salas et al. 2002; Salas et al. 2004). Consequently, it can be envisaged that almost all sub-Saharan African population groups characterised in this study will have similar genetic profiles.

Although this level of genetic homogeneity (especially of NRY) among groups will hinder the process of finding the most similar group to Palenque, the considerable increase in the number of Y chromosome markers identified in recent years (Sims et al. 2007; Karafet et al. 2008) has the potential to overcome this hurdle through the identification of NRY markers with differential frequency distributions in sub-Saharan African populations.

2.1.4. Aims

The principal aim of this study is to investigate whether genetic data are in accordance with the hypothesis of a Yombe founding group, a proposition drawn from anthropological data. A battery of NRY markers were analysed in datasets of a) Palenque, b) multiple groups from The Republic of Congo and c) peoples throughout sub-Saharan Africa. We acknowledge that while greater similarity between the Palenque and Yombe datasets than between the Palenque and any other group would strongly support the anthropological hypothesis, any alternative finding would not necessarily contradict it since there are many demographic processes that could lead to the loss from the community of NRY haplotypes contributed by the founders. We also analysed the hypervariable region (HVR-1) of mtDNA in Palenque and samples from across sub-Saharan Africa to establish whether any pattern observed in the NRY was also present in mtDNA.

Furthermore, since it is claimed that the people of Arriba are more traditional and have better conserved Palenquero than the residents of Abajo, it has been suggested that the founding men of Arriba were born in Congo, whereas Abajo was populated with slaves born in Colombia (Yves Monino, personal communication based on field work in Palenque de San Basilio). I shall also test this hypothesis by grouping samples according to their district of residence (based on sociological data obtained in the field for each sample) and examine their genetic relationship with respect to sub-Saharan African and other populations.

2.2. Materials and Methods

2.2.1. Sample Collection

Buccal swabs were collected from males over eighteen years old. The collection in Palenque de San Basilio (Bolivar, Colombia), while undertaken randomly, involved questioning potential donors to establish that no two of them shared a common paternal grandfather. Samples were collected from a very substantial proportion of individuals satisfying this criterion (estimated at >90%, n = 166; Abajo area n = 85, Arriba area n = 52). Samples collected in the Republic of Congo were made at local gatherings in different areas of Brazzaville, Pointe Noire and at the villages of Kakamoeka and Lovoulou, 90 km and 70 km inland from Pointe Noire respectively. Individuals offering to be donors were accepted randomly subject to establishing that they did not share a paternal grandfather with another donor. Buccal swabs were collected anonymously with informed consent.

Sociological data were also collected from each individual as follows: age, current residence, birthplace, self-declared cultural identity, first language, second language and religion, with similar information on the individual's father, mother, paternal grandfather and maternal grandmother. The samples were classified into groups by cultural identity. Where collections from a particular group were made in more than one location (for example, Yombe, where samples were collected in three towns: Pointe-Noire, Louvoulou and Kakamoeka), locations are represented by the mean of the latitudes and longitudes of all towns. Buccal swab samples from another 34 population groups representing West, Central-West and South-East Africa (n = 2,058), which could have potentially contributed to the Atlantic slave trade, were also included in the analysis (see Table 2.1).

Table 2.1. Details of population groups included in this study

Geographic location	Population Group (Code)	Ethnic Identity	Language	Latitude	Longitude	N
Cameroon	Bankim (BA)	Tikar	Tikar	6.083	11.483	33
	Foumban (FO)	Bamoun	Bamoun	5.729	10.902	117
	Wum (WU)	Aghem	Aghem	6.389	10.073	116
Congo	Bembe (BE)	Bembe	Bembe	-4.795	11.846	109
	Kuni (KU)	Kuni	Kuni	-4.795	11.846	68
	Lari (LA)	Lari	Lari	-4.259	15.285	62
	Mboshi (MB)	Mboshi	Mboshi	-4.259	15.285	91
	Sundi (SU)	Sundi	Sundi	-4.259	15.285	25
	Teke (TE)	Teke	Teke	-4.259	15.285	63
	Vili (VI)	Vili	Vili	-4.795	11.846	108
	Yombe (YO)	Yombe	Yombe	-4.432	12.108	65
Ghana	Asante (AS)	Asante	Akan	6.106	-1.878	94
	Ewe (EW)	Ewe	Ewe	6.6	0.467	88
	Fante (FA)	Fante	Akan	5.817	-2.817	60
Malawi	Chewa (CH)	Chewa	Nyanja	-13.607	33.918	92
	Tumbuka (TU)	Tumbuka	Tumbuka	-14.27	34.79	61
	Yao (YA)	Yao	Yao	-12.77	33.874	56
Mozambique	Sena (SE)	Multiple	Sena	-17.442	35.027	62
Nigeria	Abak (AB)	Annang	Annang	5.05	7.717	56
	Afaha Eket (AE)	Ibibio	Ibibio	4.717	7.867	48
	Afaha Okpo (AO)	Oron	Oron	4.833	8.233	50
	Afaha Ukwong (AU)	Oron	Oron	4.75	8.25	49
	Awa-Onna (AW)	Ibibio	Ibibio	4.69	7.815	28
	Calabar (CA)	Igbo	Igbo	4.95	8.317	99
	Ediene Ikono (ED)	Ibibio	Ibibio	4.783	7.883	48
	Efut Akpabuyo (EF)	Efik	Efik	4.908	8.442	48
	Efut Odukpani (EO)	Efik	Efik	5.167	7.983	49
	Ejagham-Akamkpa (EA)	Ekoi	Ejagham	5.35	8.35	47
	Ejagham-Calabar (EC)	Ekoi	Ejagham	4.95	8.317	83

	Eziagu Nenwe (EZ)	Igbo	Igbo	6.117	7.517	49
	Ikono (IK)	Annang	Annang	4.992	7.758	42
	Itam (IT)	Ibibio	Ibibio	5.042	7.842	50
	Nike Enugu (NE)	Igbo	Igbo	6.433	7.483	54
	Nnung Ndem-Onna (NN)	Ibibio	Ibibio	4.633	7.85	48
	Nsit (NS)	Ibibio	Ibibio	4.833	7.9	36
	Ntan Ibiono (NT)	Ibibio	Ibibio	5.233	7.933	50
	Obong Itam (OB)	Ibibio	Ibibio	5.133	7.967	50
	Oku-Itu (OI)	Ibibio	Ibibio	5.133	7.933	49
	Oku-Uyo (OU)	Ibibio	Ibibio	5.1	7.967	48
	Ukpom Ete (UE)	Ibibio	Ibibio	4.62	7.65	50
	Uwanse (UW)	Efik	Efik	4.95	8.317	50
South Africa	Bantu-Pretoria (BN)	Bantu	Bantu	-25.746	28.187	98

2.2.2. DNA Extraction

DNA from all Congolese and Palenque samples was extracted using the Gentra protein precipitation method (Gentra Systems, Minneapolis) (see Appendix A). Previously collected buccal swab DNA samples from ethnic groups across sub-Saharan Africa included in this study were extracted by standard phenol-chloroform method. The range and mean of sample sizes of these ethnic groups are 25-118 and 62 respectively.

2.2.3. Y-chromosome typing

A combination of Unique Event Polymorphisms (UEP) and short tandem repeats (STRs) in the paternally inherited NRY were typed in all samples from Palenque (n=166) and the eight Congolese ethnic groups (n=597). The polymorphic markers are six STRs (DYS19, DYS388, DYS390, DYS391, DYS392 and DYS393) and four UEPs (M191, U175, U290 and U181) characterising the E1b1a haplogroup, which is the modal haplogroup throughout the area of the expansion of the Bantu-speaking peoples. STR repeat sizes were assigned according to the nomenclature of Kayser et al. (1997). The four UEPs were typed using a tetra primer ARMS PCR method (Ye et al. 2001) with minor modifications.

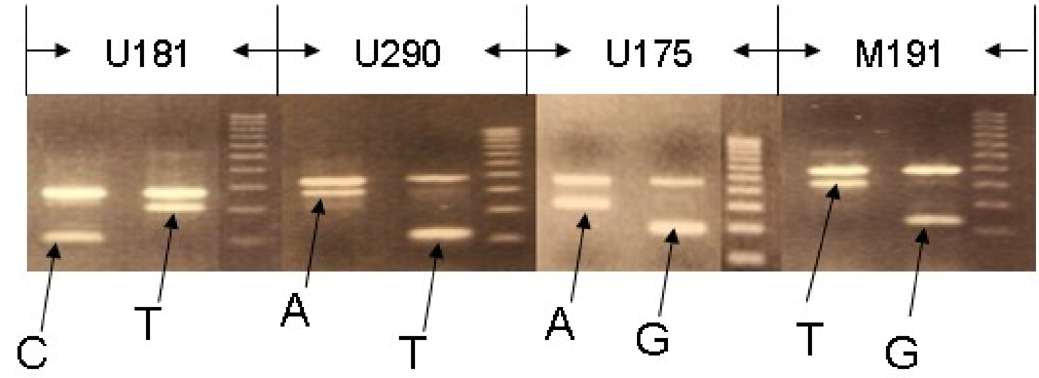
The outer and two inner fragments were amplified in a 10-µl reaction volume containing 1 µl (~ 1 ng) of template DNA, 1.6 µl (50 µM) dNTPs, 9.3 nM TaqStart monoclonal antibody (BD Biosciences Clontech, Oxford, UK), 1 µl of 10x Taq buffer and 0.13 units of Taq DNA polymerase (HT Biotech, Cambridge, UK) and outer and inner primers (see Table 2.2 for primer details). All samples (96-well plates) were then placed on a thermocycler under the following conditions: denaturation at 95ºC for 5 min, followed by 35 cycles of denaturation (95ºC) for 45 s, annealing (see Table 2.2 for annealing temperatures) for 45 s and elongation (72ºC) for 45 s.

Table 2.2. Details of the primers used in the tetra-primer ARMS PCR reactions

UEP	Primer name	Primer sequence (5'-3')	Melting temperature (°C)	Fragment size (bp)	Primer concentration (µM)	PCR annealing temperature (°C)
M191 T>G	Outer forward primer	AATACCAGGCCGACATGGCAGCTA	64.4	399	0.15	60.5
	Outer reverse primer	CTACAAGCACGTACCACAGCGCCA	66.1		0.15	
	Inner forward primer	CATTTTTTTCTTTACAACTTGACCAG	56.9	133	0.75	
	Inner reverse primer	CACACCAAAATATCTCATATTTTCGTA	57.4	318	0.75	
U175 G>A	Outer forward primer	CCTTTAACACACTTCACAACATGG	59.3	274	0.3	53.0
	Outer reverse primer	GTGTCACTTTTCATTGTCTGG	55.9		0.3	
	Inner forward primer	CCACAGGTGCTAATGAAATCG	57.9	106	0.3	
	Inner reverse primer	ATGACCAGGAGAAGTCAAAAT	54.0	209	0.3	
U290 T>A	Outer forward primer	GCTATTGGAGAGCCTCGCTGTG	61.0	449	0.15	61.0
	Outer reverse primer	AGGAAGCAATTTTCCTACCTGCCA	59.5		0.15	
	Inner forward primer	GATAGGTGTGGGAATTGATGGCATT	58.3	193	0.75	
	Inner reverse primer	GATGGCCATCAGTCCCCAGT	60.4	300	0.75	
U181 C>T	Outer forward primer	GGTCTAGTGCACAGTGGTATCCA	57.1	389	0.15	62.0
	Outer reverse primer	AGAGCTCTCTCAAATCTGTGTTGG	57.5		0.15	
	Inner forward primer	AGTGTCTTTGTTTTGGCAAGAAC	61.8	123	0.75	
	Inner reverse primer	CTACCCTTGTATCAGAATACAGTTCTTA	61.7	316	0.75	

The final step of the PCR program was a 7 min extension at 72ºC before a 30 min hold at 4ºC. Figure 2.2 shows examples of agarose gels for both alleles at all four E1b1a UEP.

Figure 2.2. Scan of an agarose gel showing the result of tetra-primer ARMS PCR for the detection of E1b1a UEP alleles.



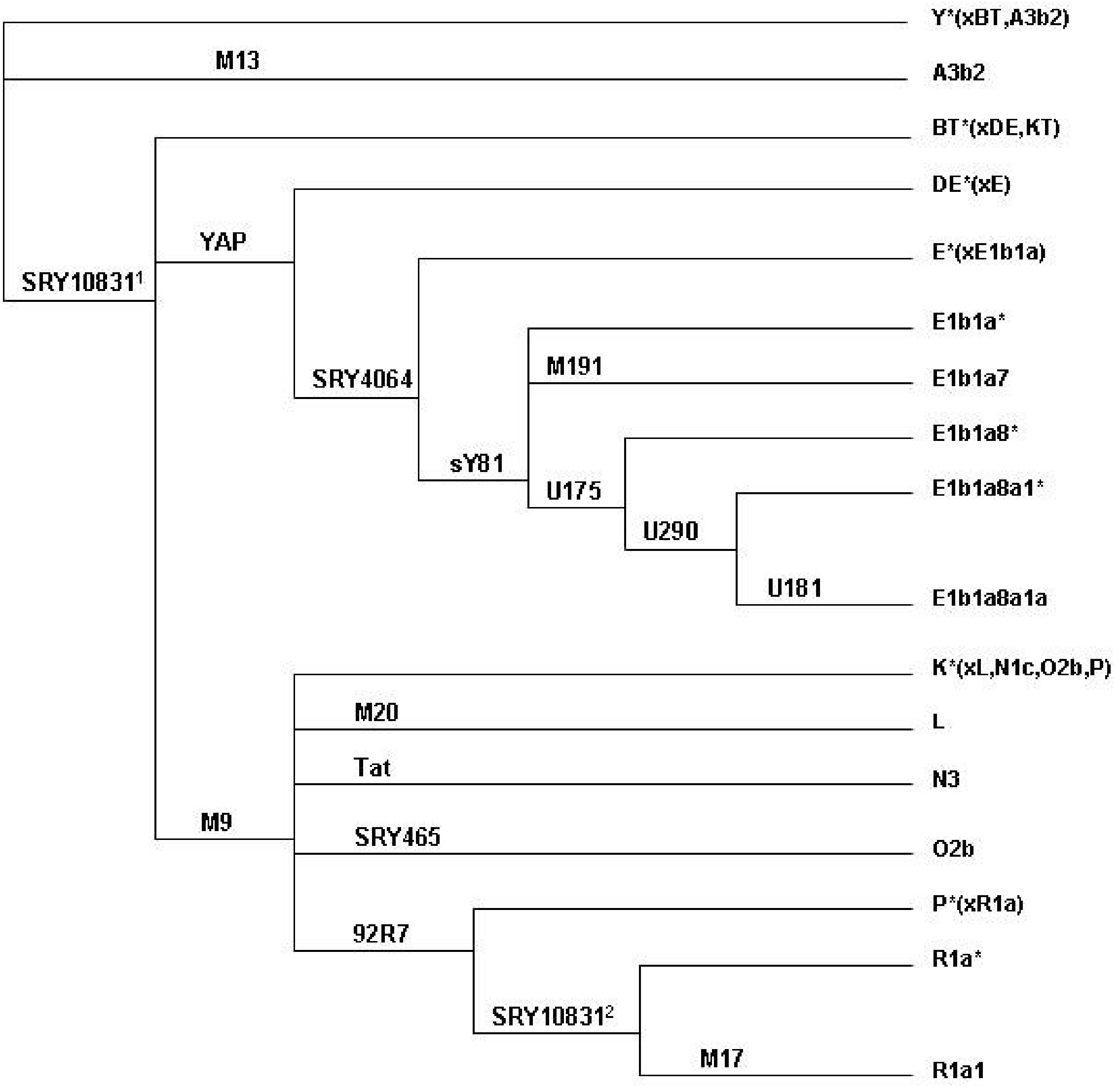
Note: The common top PCR band in all samples is the outer (control) fragment. In all cases, the molecular DNA marker has 10 bands of sizes 1000 to 100 bp from top to bottom. For exact sizes of each fragment, see Table 2.2.

Where samples were not E1b1a (no derived alleles at the 4 UEP markers), a further six to eleven UEPs (TCGA UEP1 and UEP2 kits: sY81, SRY4064, YAP, SRY10831, M13, M9, SRY465, M20, Tat, 92R7 & M17) were typed (Thomas et al. 1999). The four E1b1a-specific UEPs were also typed for E1b1a samples (previously characterised) from the 34 sub-Saharan ethnic groups (n=2,014) detailed in Table 2.1. NRY haplogroups were classified according to the nomenclature of the Y Chromosome Consortium (Karafet et al. 2008) (see Figure 2.3).

2.2.4. mtDNA typing

The mtDNA HVR-1 region of all Congolese groups and Palenque was sequenced as described by Thomas et al. (2002), but with the following modification: primers conL1-mod, conL2 and conH3 were replaced by conL849 (CTA TCT CCC TAA TTG AAA ACA AAA TA), conL884 (TGT CCT TGT AGT ATA A) and conHmt3 (CCA GAT GTC GGA TAC AGT TC) respectively for better sequencing results. For all samples, HVR-1 Variable Site Only (VSO) haplotypes were determined by comparing generated HVR-1 sequences of nucleotide range 16020-16400 with the Cambridge Reference Sequence (Anderson et al. 1981). It should be noted that Andrews et al. (1999) re-evaluated the mtDNA reference sequence by re-sequencing the entire mtDNA genome and found no discrepancies in the HVR-1 region and therefore validated the use of the old sequence for the purposes of this study.

Figure 2.3. Genealogical relationships of UEP markers used to define NRY haplogroups



Note: E1b1a* is the E1b1a clade excluding E1b1a7 and E1b1a8

VSO haplotypes were defined by the type of occurring mutations (substitution, insertion or deletion) and their corresponding nucleotide positions. In order to extend the dataset, HVR-1 Variable Site Only (VSO) haplotypes were also determined for population groups previously typed for the HVR-1 region (n = 30; all sub-Saharan groups except Sena, Tumbuka, Bantu speakers-Pretoria and Yao). As the nucleotide range of these samples for the HVR-1 region was 16023-16380, the HVR-1 coverage in Congolese samples and Palenque was reduced to this range for comparisons with these groups.

Chromatograms of all samples were inspected visually. If no significant trace of noise was found, the ends of the sequence were then trimmed down to the standard nucleotide range (16020-16400) by visually inspecting the sequence. A contig was then formed for each set of 96 samples (i.e. one sequencing run) and each mutation site was examined visually to determine whether it was a SNP, insertion or deletion. All samples with any ambiguous sites were resequenced.

2.2.5. Statistical analysis

Gene diversity and its standard error were estimated from unbiased formulae of Nei (1987). Mean number of pairwise nucleotide differences were also calculated for mtDNA HVR-1 sequences in each group. The genetic distances used were a) F_{ST} (Reynolds, Weir & Cockerham 1983) (based on UEP haplogroups, UEP+STR haplotypes and mtDNA HVR-1 VSO haplotypes), b) R_{ST} (Slatkin 1995) (based on six STR on the NRY) and c) the Kimura's two-parameter model with gamma distribution of value 0.47 (Kimura 1980) (on mtDNA HVR-1 sequences). Significance of genetic distances was assessed by permutation test, where a null distribution is formed by calculating all possible values of the genetic distance under rearrangements of haplogroups/haplotypes in the two observed datasets. The permutation test was repeated over 1,000 times to give a null distribution. Pairwise genetic differences between population groups were also estimated using the Exact Test of Population Differentiation (ETPD) (Raymond & Rousset 1995), which is analogous to the Fisher's Exact test with an m x n contingency table, where m is the number of groups and n is the number of haplotypes. All the above was performed using Arlequin software version 3.0 (Excoffier et al 2005).

Principal Component Analysis (PCA) was performed using the ‘R’ environment of statistical computing (www.R-project.org) by implementing the ‘princomp’ function based on E1b1a component haplogroup frequencies and visualised using the ‘plot’ function. PCA plots were used to visualise relationships among groups. When a candidate population group for the origin of Palenque, which has the most similar NRY distribution, did not have a significant difference from another population group in terms of their genetic distance (FST), the significance of difference between the two groups with respect to Palenque was assessed by comparing FST distributions of 100,000 bootstrap re-sampled datasets. All bootstrap re-sampling was performed in the statistical package software R (2.4.0).

2.3. Results

2.3.1. Frequencies of NRY haplotypes and NRY-based genetic distances

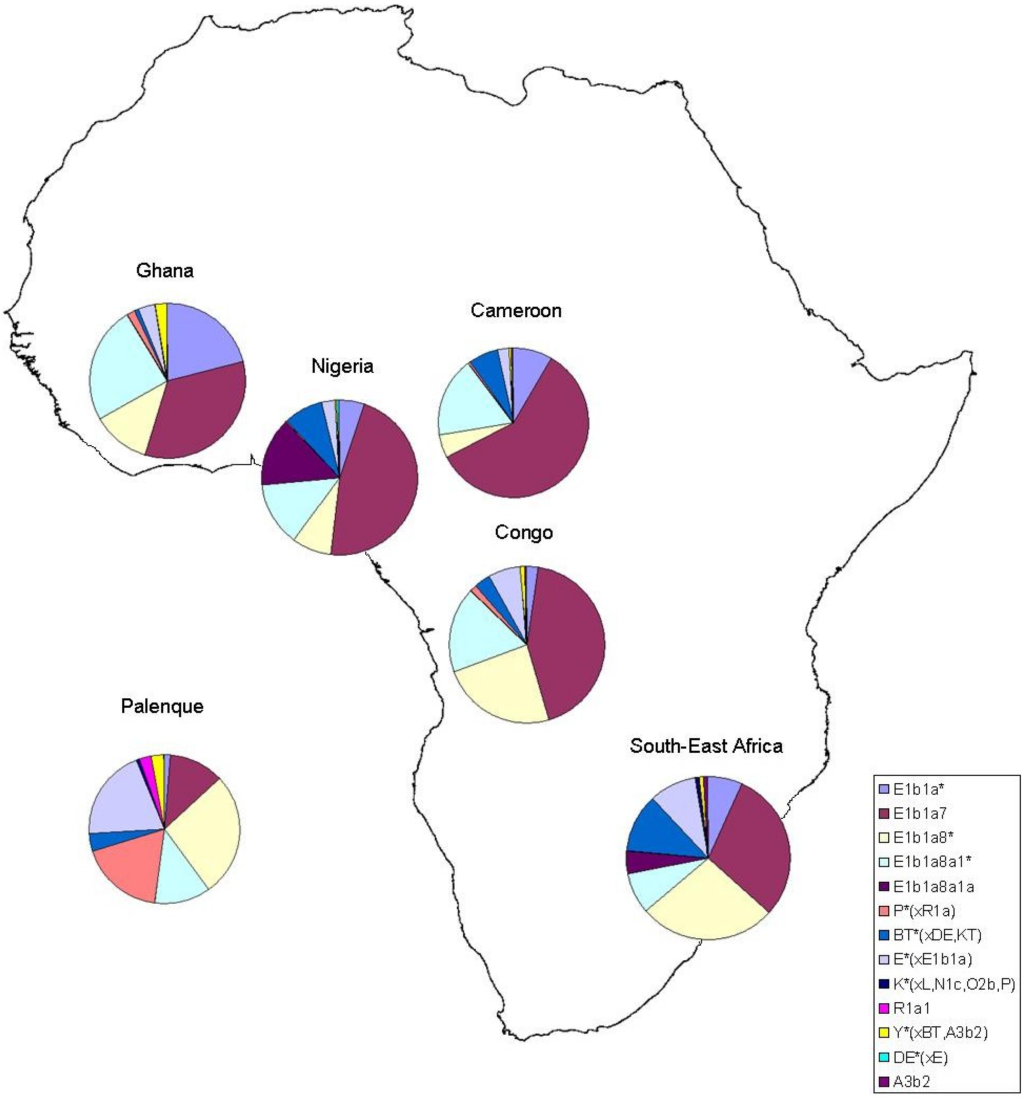
The frequencies of all NRY haplogroups in the Palenque and in 42 sub-Saharan groups analysed in this study are included in Table 2.3. The phylogenetic relationships of the haplogroups and a visual representation of the distribution of haplogroups can be seen in Figures 2.3 and 2.4 respectively.

Table 2.3. Haplogroup frequencies in Palenque and 42 sub-Saharan African population groups from seven broad geographic regions. Modal NRY haplogroup shown in bold type.

NRY UEP Haplogroup (according to the nomenclature proposed by Karafet et al. (2008))	**E1b1a***	**E1b1a7**	**E1b1a8***	**E1b1a8a1***	**E1b1a8a1a**	**P*(xR1a)**	**BT*(xDE,KT)**	**E*(xE1b1a)**	**K*(xL,N1c,O2b,P)**	**R1a1**	**Y*(xBT,A3b2)**	**DE*(xE)**	**A3b2**
Palenque-Abajo (PAB)	0.012	0.129	**0.341**	0.071	0.000	0.153	0.012	0.212	0.000	0.047	0.024	0.000	0.000
Palenque-Arriba (PAR)	0.019	0.096	0.173	**0.231**	0.000	0.192	0.096	0.154	0.019	0.000	0.019	0.000	0.000
Palenque total (PA)	0.013	0.120	**0.267**	0.120	0.000	0.180	0.040	0.200	0.007	0.027	0.027	0.000	0.000
Bembe (BE)	0.037	**0.321**	0.211	0.284	0.000	0.018	0.000	0.110	0.000	0.000	0.000	0.000	0.018
Kuni (KU)	0.015	**0.515**	0.162	0.176	0.000	0.000	0.044	0.074	0.000	0.000	0.015	0.000	0.000
Lari (LA)	0.000	**0.484**	0.177	0.194	0.000	0.000	0.032	0.113	0.000	0.000	0.000	0.000	0.000
Mboshi (MB)	0.000	**0.440**	0.286	0.187	0.000	0.011	0.044	0.000	0.000	0.000	0.033	0.000	0.000
Sundi (SU)	0.040	**0.600**	0.200	0.040	0.000	0.000	0.040	0.080	0.000	0.000	0.000	0.000	0.000
Teke (TE)	0.000	**0.429**	0.222	0.159	0.000	0.032	0.063	0.079	0.000	0.000	0.016	0.000	0.000

Vili (VI)	0.056	**0.500**	0.250	0.139	0.000	0.009	0.028	0.019	0.000	0.000	0.000	0.000	0.000
Yombe (YO)	0.031	0.277	**0.369**	0.138	0.000	0.015	0.046	0.092	0.000	0.000	0.031	0.000	0.000
Congo Total	0.024	**0.430**	0.239	0.181	0.000	0.012	0.034	0.066	0.000	0.000	0.012	0.000	0.003
Bankim (BA)	**0.364**	0.182	0.152	0.152	0.000	0.030	0.061	0.061	0.000	0.000	0.000	0.000	0.000
Foumban (FO)	0.068	**0.667**	0.068	0.051	0.000	0.000	0.085	0.043	0.000	0.000	0.009	0.000	0.009
Wum (WU)	0.026	**0.621**	0.000	0.310	0.000	0.000	0.043	0.000	0.000	0.000	0.000	0.000	0.000
Cameroon Total	0.086	**0.586**	0.049	0.177	0.000	0.004	0.064	0.026	0.000	0.000	0.004	0.000	0.004
Asante (AS)	0.245	**0.309**	0.106	0.245	0.000	0.000	0.000	0.043	0.000	0.000	0.053	0.000	0.000
Ewe (EW)	0.114	**0.420**	0.102	0.273	0.000	0.023	0.023	0.045	0.000	0.000	0.000	0.000	0.000
Fante (FA)	0.300	**0.250**	0.167	0.217	0.000	0.033	0.000	0.017	0.000	0.000	0.017	0.000	0.000
Ghana Total	0.211	**0.335**	0.120	0.248	0.000	0.017	0.008	0.037	0.000	0.000	0.025	0.000	0.000
Chewa (CH)	0.043	0.239	**0.293**	0.130	0.054	0.000	0.076	0.130	0.011	0.000	0.000	0.000	0.022
Tumbuka (TU)	0.033	**0.410**	0.230	0.082	0.033	0.000	0.082	0.098	0.016	0.000	0.000	0.000	0.016
Yao (YA)	0.071	**0.321**	0.250	0.054	0.036	0.000	0.196	0.071	0.000	0.000	0.000	0.000	0.000
Malawi Total	0.048	**0.311**	0.263	0.096	0.043	0.000	0.110	0.105	0.010	0.000	0.000	0.000	0.014
Abak (AB)	0.054	**0.357**	0.107	0.161	0.232	0.000	0.054	0.036	0.000	0.000	0.000	0.000	0.000
Afaha Eket (AE)	0.104	**0.354**	0.042	0.125	0.208	0.000	0.125	0.042	0.000	0.000	0.000	0.000	0.000
Afaha Okpo (AO)	0.020	**0.460**	0.060	0.080	0.280	0.000	0.020	0.060	0.000	0.000	0.000	0.020	0.000
Afaha Ukwong (AU)	0.041	**0.408**	0.102	0.143	0.163	0.000	0.102	0.020	0.000	0.000	0.020	0.000	0.000
Awa-Onna (AW)	0.000	**0.500**	0.000	0.179	0.107	0.000	0.179	0.036	0.000	0.000	0.000	0.000	0.000
Calabar (CA)	0.061	**0.475**	0.121	0.101	0.141	0.000	0.071	0.010	0.000	0.000	0.010	0.010	0.000
Ediene Ikono (ED)	0.083	**0.479**	0.042	0.083	0.125	0.000	0.125	0.063	0.000	0.000	0.000	0.000	0.000
Efut Akpabuyo (EF)	0.125	**0.438**	0.042	0.042	0.125	0.021	0.188	0.021	0.000	0.000	0.000	0.000	0.000
Efut Odukpani (EO)	0.020	**0.449**	0.143	0.082	0.204	0.000	0.061	0.041	0.000	0.000	0.000	0.000	0.000
Ejagham-Akamkpa (EA)	0.021	**0.426**	0.085	0.277	0.043	0.021	0.085	0.043	0.000	0.000	0.000	0.000	0.000
Ejagham-Calabar (EC)	0.024	**0.518**	0.048	0.193	0.108	0.024	0.060	0.024	0.000	0.000	0.000	0.000	0.000
Eziagu Nenwe (EZ)	0.286	**0.469**	0.020	0.102	0.082	0.000	0.041	0.000	0.000	0.000	0.000	0.000	0.000
Ikono (IK)	0.000	**0.429**	0.167	0.143	0.071	0.000	0.167	0.024	0.000	0.000	0.000	0.000	0.000
Itam (IT)	0.080	**0.520**	0.060	0.120	0.160	0.000	0.040	0.000	0.000	0.000	0.020	0.000	0.000
Nike Enugu (NE)	0.019	**0.685**	0.000	0.074	0.130	0.000	0.056	0.019	0.000	0.000	0.000	0.019	0.000
Nnung Ndem-Onna (NN)	0.042	**0.563**	0.063	0.250	0.021	0.000	0.021	0.021	0.000	0.000	0.000	0.021	0.000
Nsit (NS)	0.028	**0.361**	0.111	0.250	0.167	0.000	0.056	0.028	0.000	0.000	0.000	0.000	0.000
Ntan Ibiono (NT)	0.040	**0.380**	0.160	0.140	0.180	0.000	0.080	0.020	0.000	0.000	0.000	0.000	0.000
Obong Itam (OB)	0.000	**0.480**	0.060	0.140	0.120	0.020	0.100	0.040	0.020	0.000	0.000	0.020	0.000
Oku-Itu (OI)	0.000	**0.510**	0.102	0.102	0.163	0.000	0.122	0.000	0.000	0.000	0.000	0.000	0.000
Oku-Uyo (OU)	0.042	**0.396**	0.083	0.208	0.167	0.000	0.042	0.063	0.000	0.000	0.000	0.000	0.000
Ukpom Ete (UE)	0.060	**0.480**	0.140	0.100	0.080	0.000	0.080	0.040	0.000	0.000	0.000	0.020	0.000
Uwanse (UW)	0.060	**0.460**	0.060	0.060	0.240	0.000	0.100	0.000	0.000	0.000	0.020	0.000	0.000
Nigeria Total	0.053	**0.464**	0.080	0.135	0.145	0.004	0.082	0.027	0.001	0.000	0.003	0.005	0.000
Bantu-speakers Pretoria (BN)	0.061	0.255	**0.286**	0.071	0.082	0.000	0.153	0.051	0.010	0.000	0.031	0.000	0.000
Sena-Mozambique (SE)	0.161	**0.306**	0.274	0.032	0.016	0.000	0.081	0.129	0.000	0.000	0.000	0.000	0.000
Grand Total	0.064	**0.417**	0.149	0.150	0.068	0.016	0.066	0.054	0.002	0.001	0.009	0.002	0.002

Figure 2.4. Visual representation of haplogroup distributions in sub-Saharan African populations and Palenque.



Note: South-East Africa is represented by population groups CH, TU, YA, SE and BN.

The 15 typed UEP makers defined 13 observed NRY haplogroups, of which ten were observed in Palenque. The modal haplogroup in the Palenque was E1b1a8* (27%), but the two moieties of the village had different modal haplogroups: Abajo, E1b1a8* (34%); Arriba, E1b1a8a1* (23%); using the nomenclature of the Y-chromosome Consortium (Karafet et al. 2008). Gene diversity in Palenque, based on UEP haplogroups, was 0.830 ± 0.013, whereas the equivalent statistic in the sub-Saharan African dataset (n = 2,649) was 0.753 ± 0.007 and in each individual group was lower than in Palenque, ranging from 0.514-0.820. Gene diversity for each broad geographic region is included in Table 2.4.

Table 2.4. Gene diversity measures based on UEP haplogroups and UEP+STR haplotypes in five broad geographic regions in sub-Saharan Africa.

Geographic region	NRY Marker Set	Gene Diversity ± SD	Gene Diversity Range among groups	Gene Diversity Mean (Variance)	N
Cameroon (n=3)	UEP	0.613 ± 0.029	0.520-0.805	0.621 (0.025)	266
	UEP+STR	0.954 ± 0.005	0.893-0.962	0.924 (0.001)	
Congo (n=8)	UEP	0.721 ± 0.012	0.613-0.767	0.703 (0.003)	591
	UEP+STR	0.951 ± 0.003	0.928-0.957	0.941 (0.0001)	
Ghana (n=3)	UEP	0.769 ± 0.012	0.731-0.784	0.764 (0.001)	242
	UEP+STR	0.968 ± 0.004	0.962-0.974	0.968 (0.00004)	
Nigeria (n=23)	UEP	0.729 ± 0.011	0.514-0.802	0.724 (0.003)	1181
	UEP+STR	0.969 ± 0.002	0.936-0.983	0.967 (0.0001)	
South East Africa (n=5)*	UEP	0.805 ± 0.011	0.766-0.820	0.799 (0.0004)	369
	UEP+STR	0.971 ± 0.003	0.964-0.976	0.969 (0.00003)	

*This region includes the following groups: CH, TU, YA, SE and BN. N, n and SD are total sample size, number of groups and standard deviation respectively.

Haplotype (UEP+STR) sharing across groups is shown in Supplementary Table 2.S5, where it can be seen that the number and proportion of haplotypes shared between the Palenque and each other group ranges from 5-12 (0.111-0.267) and 7-68 (0.047-0.453).

Genetic distances between Palenque and the sub-Saharan African groups at the NRY UEP characterisation are provided in Supplementary Table 2.S1, where it can be seen that Chewa has the smallest Fst with Palenque and is not statistically significant. The second most similar group to Palenque was the Yombe, which had a non-significant Fst with Chewa (P = 0.595).

The NRY E1b1a clade comprised 52.0% of the Palenque and 86.6% of the 42 sub-Saharan African groups (range of 73.2-95.9%, mean of 86.3% and variance of 0.3%) (see

Table 2.5). Restricting NRY UEP analysis just to the component haplogroups of this clade (NRY UEP-E1b1a) showed the genetic distance between Palenque and Yombe to be the smallest (genetic distances are included Supplementary Table 2.S2).

Table 2.5. E1b1a clade proportions in five broad geographic regions in sub-Saharan Africa. Notations are the same as in Table 2.4.

Geographic region	E1b1a frequency	E1b1a frequency Range among groups	E1b1a frequency Mean (Variance)	N
Cameroon (n=3)	0.898 ± 0.019	0.848-0.957	0.887 (0.004)	266
Congo (n=8)	0.873 ± 0.014	0.810-0.944	0.867 (0.002)	591
Ghana (n=3)	0.913 ± 0.018	0.904-0.933	0.916 (0.0002)	242
Nigeria (n=23)	0.887 ± 0.010	0.771-0.959	0.874 (0.003)	1181
South East Africa (n=5)	0.764 ± 0.022	0.732-0.790	0.765 (0.0006)	369

2.3.2. Comparisons of NRY datasets

Supplementary Tables 2.S3 and 2.S4 show the results of pairwise comparisons using the Exact Test of Population Differentiation (ETPD) between all groups based on all NRY haplogroups and just the E1b1a component haplogroups. At both levels, the Cameroonian groups (Bankim, Foumban and Wum) along with Palenque showed the highest level of differentiation compared with all other groups, with ≥ 40 out of total of 42 significant differences observed. Notably, pairwise ETPD test of Palenque and all 42 sub-Saharan African groups (including the eight Congolese groups) were significant at 0.1% and 5% significance levels, based on UEP and UEP-E1b1a, except for Yombe (UEP P = 0.0011, UEP-E1b1a P = 0.504).

2.3.3 Yombe and Chewa

The two groups most similar to Palenque were not statistically distinguishable from each other at either UEP or E1b1a-UEP levels, using ETPD (P ≥ 0.318). Because a relatively higher level of differentiation was observed at the UEP-E1b1a level, genetic distance was recalculated at this level between Yombe-Palenque (YP) and Chewa-Palenque (CP) using Fst with bootstrap re-sampling of 100,000 repeats. 95% confidence intervals (the middle 95% of estimated Fst values) of YP and CP pairwise Fst (see Table 2.6) showed no indication of a closer relationship between Yombe and Palenque. However, estimates of mean and mode were lower in the YP distribution than in the CP distribution.

Table 2.6. Summary statistics of bootstrap resampled Fst distributions based on E1b1a NRY clade using two definitions of Fst. Negative Fst results are artefacts obtained because of the computation method used. CI stands for confidence interval.

Fst definition	Pairwise comparison	95% CI	Mean	Mode
Wright's Fst	Yombe-Palenque (YP)	0.0009 - 0.0476	0.0150	0.0035
	Chewa-Palenque (CP)	0.0032 - 0.0504	0.0190	0.0115
Reynolds' Fst (Theta)	Yombe-Palenque (YP)	-0.0061 - 0.0867	0.0225	0.0025
	Chewa-Palenque (CP)	-0.0004 - 0.0900	0.0305	0.0165

2.3.4 Principal Component Analysis

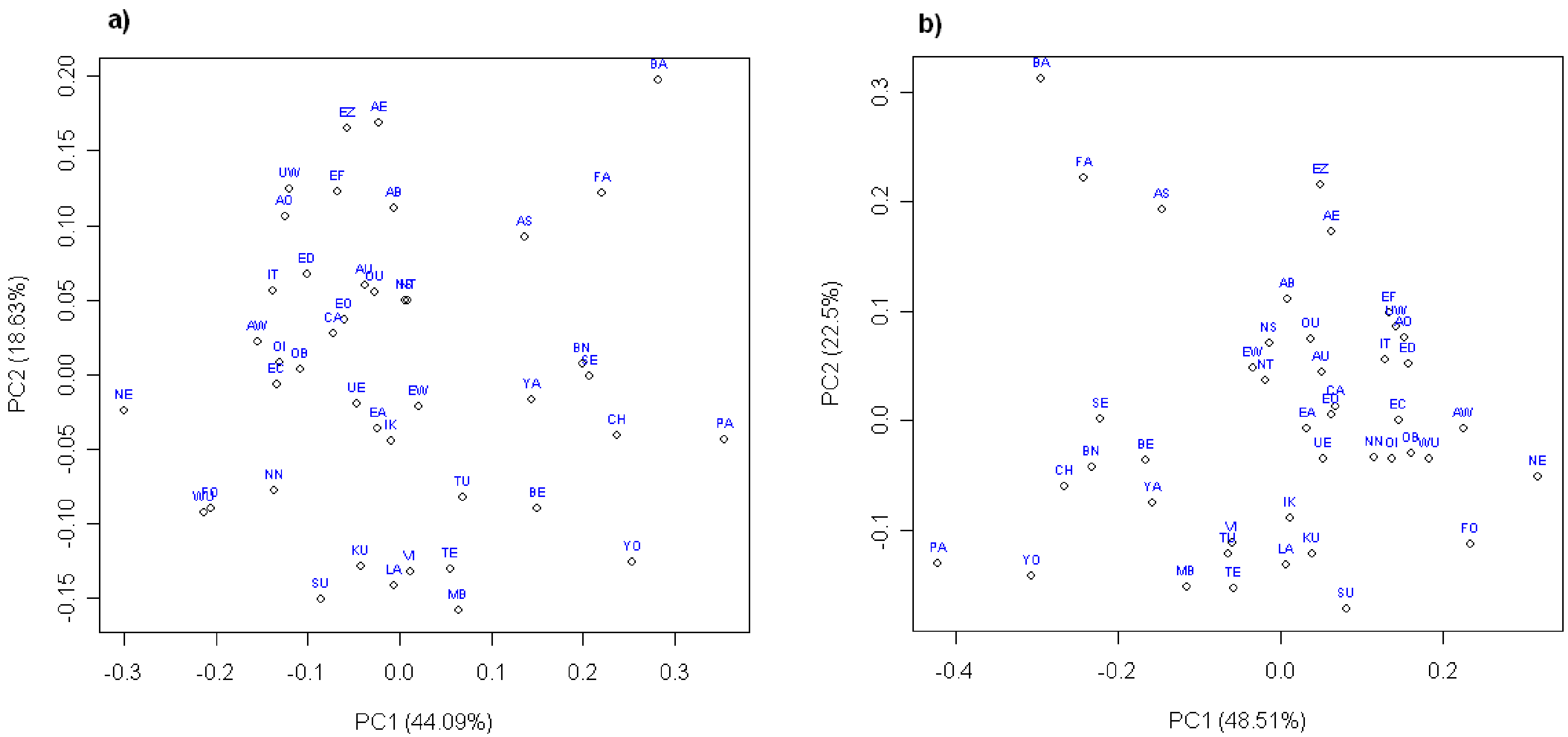
Principal component analysis was used to visualise patterns of genetic relationships among groups. Among the 42 groups Yombe and Chewa relatively had the same level of similarity to Palenque at the UEP level. However, based on UEP-E1b1a, the most similar group to Palenque was Yombe (see Figure 2.5).

2.3.5. The distribution of mtDNA variation

The Palenque dataset contained 26 VSO haplotypes, based on sequencing 360 nucleotides of the HVR-1 region of the mtDNA genome. The modal haplotype was at a frequency of 0.166 and five common haplotypes accounted, together, for 66.2% of the total. Nei's gene diversity, based on VSO haplotypes, was 0.9031 ± 0.010.

Across all sub-Saharan African groups (38 groups), 723 VSO haplotypes (427 singletons) were observed. The modal haplotype frequency ranged from 0.040-0.147, where Obong Itam and Bankim had the lowest and highest frequency respectively. Gene diversity for the combined set was 0.9933 ± 0.004 and in individual groups ranged from 0.968 in Bankim to 0.997 in two Nigerian groups (Ejagham-Akamkpa and Obong Itam), with mean of 0.987 and variance of 0.00005.

Figure 2.5. Visual representation of genetic relationships among all groups using PCA based on a) NRY UEP and b) NRY UEP-E1b1a.



Percentages in parentheses are the amount of variation explained by each component.

The frequencies of all mtDNA haplotypes are included in Supplementary Table 2.S6.

2.3.6. Comparison of mtDNA across datasets

Comparison of all 38 sub-Saharan groups, using Fst, showed 457 out of 703 (65%) to be significant at the 5% level. Individual groups had significant genetic distances with 10-37 other groups. (Ukpom Ete had only 10 significant genetic distances, whereas Bankim, Bembe and Chewa had 37 each). Fst between Palenque and the sub-Saharan African groups were all significant ($P < 0.0001$) and were in the range of 0.037-0.066, where Oku-Uyo from Nigeria and Chewa had the lowest and highest genetic distances.

Based on VSO haplotypes, all pairwise ETPD between Palenque and 42 sub-Saharan African groups (including the eight Congolese groups) were significant ($P < 0.0001$).

2.3.7. Intra-village analysis of Palenque

Summary statistics were calculated for both districts (Abajo and Arriba) in Palenque. The modal NRY haplogroups in Abajo and Arriba were E1b1a8* (34.1%) and E1b1a8a1* (23.1%) respectively (Table 2.3). The proportion of the E1b1a clade in Abajo and Arriba was 55.3% and 51.9% respectively (not significantly different). Gene diversity based on all UEP haplogroups in Abajo and Arriba was 0.8000 ± 0.0242 and 0.8529 ± 0.0188 respectively. Restricting analysis to E1b1a NRY types, gene diversity in Abajo and Arriba was markedly reduced to 0.5597 ± 0.0632 and 0.6809 ± 0.0495 respectively. At the UEP+STR level the modal haplotype was one step away from the EBSP modal haplotype (i.e. E1b1a-15-12-21-10-11-13) in Abajo (E1b1a-16-12-21-10-11-13; 10.6%) and in Arriba (E1b1a-15-12-21-10-12-13; 19.2%), whereas the EBSP modal haplotype was at a frequency of 5.9% and 5.8% in Abajo and Arriba respectively.

To investigate the hypothesis, derived from observed cultural differences and local opinion (Yves Monino, personal field notes), that the village areas of Abajo and Arriba can be distinguished from each other several statistical tests were performed involving, separately, both the NRY and mtDNA genetic systems. Initially, a set of different genetic distance estimators were calculated based on NRY and mtDNA markers (see Table 2.7).

Table 2.7. Genetic distances between Abajo and Arriba based on NRY and mtDNA markers. Significant p-values at 5% level are shown in bold type.

Genetic system	Marker/s	Distance estimator	Distance	P-value
NRY	UEP ALL	Fst	0.027	**0.018**
	UEP E1b1a only		0.116	**0.027**
	UEP+STR ALL		0.019	**<0.00001**
	UEP+STR E1b1a only		0.041	**0.018**
	STR	Rst	0.023	**0.045**
mtDNA	HVR-1 VSO	Fst	<0.001	0.982
	HVR-1 sequence	K2P	<0.001	0.712

Strikingly, genetic distances at the mtDNA level were not significant, whereas all distance measures based on NRY markers were significant at the 5% level.

Genetic difference of Abajo and Arriba was also assessed by ETPD. The NRY genetic system showed significant difference between Abajo and Arriba at the 5% level (UEP, UEP-E1b1a and UEP+STR (see Table 2.7)), however, no significant genetic difference was found between the two districts based on mtDNA HVR-1 haplotypes (Fst) or sequences (π) ($p > 0.98$).

2.3.8 Abajo and Arriba in the context of sub-Saharan Africa

Genetic differences between each area of the village and the sub-Saharan Africa groups were independently assessed by the ETPD, based on NRY UEP and UEP E1b1a component haplogroups. Analysing UEP, both Abajo and Arriba were significantly different from all the sub-Saharan groups. Notably, analysing UEP haplogroups in the E1b1a clade, all groups were also significantly different from Abajo at the 5% level, except Yombe ($p = 0.454$) and Chewa ($p = 0.082$), whereas three groups could not be distinguished from the Arriba (Bembe ($p = 0.272$), Chewa ($p = 0.069$), Yombe ($p = 0.063$)).

Genetic distances between Abajo and Arriba, as separate datasets, and 42 population groups in sub-Saharan Africa were also calculated based on NRY UEP and UEP E1b1a component haplotypes, but not VSO haplotypes, since no significant genetic distance was observed between the two districts, based on mtDNA (see Table 2.7). Analysing UEP, Chewa were the closest to Abajo, whereas based on E1b1a component haplotypes, Yombe were. In the case of Arriba, neither Yombe nor Chewa were the closest.

2.4. Discussion

2.4.1 Answering the principal questions of the study

There were three objectives of this study as enunciated in the introduction to this chapter:

1) are genetic data consistent with a Yombe-speaking paternal founding group of the Palenque?
2) is there any pattern in the distribution of maternal ancestry based on mtDNA HVR-1 haplotypes similar to that displayed by analysis of the NRY?
3) can the two groups of residents of Palenque de San Basilio, occupying Abajo and Arriba respectively, be distinguished from each other and, if so, are the founding residents of Arriba more likely to have been born in the Congo than were the founding residents of Abajo?

It is first necessary to consider which set of NRY and mtDNA markers are appropriate to provide the power necessary to undertake analyses. On the NRY there are the following choices: UEP haplogroups, UEP+STR haplotypes characterised by both SNPs and STRs, UEP haplogroups comprising the E1b1a clade and STRs without SNPs. With respect to mtDNA there are VSO haplotypes available which could be used with or without imputing haplogroups based on the haplotype and analysed either as haplotypes or on the basis of nucleotide diversity.

The choices were made for the reasons set out below:

UEP haplotypes: these were analysed because their frequencies have been shown to vary among groups in ways that reflect common origin or gene flow (Underhill et al. 2001). A problem in addressing the question posed in this study concerning the possible Yombe origin of the Palenque is that the Palenque are anticipated to have experienced extensive inward NRY gene flow from non-African populations, which the African groups have not, and that this, while not invalidating the adopted approach, might make interpretation of genetic distances calculated using UEP haplotypes more difficult.

E1b1a clade component haplotypes: analysis of these NRY have the benefits associated with the UEP haplotypes method above with an added advantage that NRY likely to be due to non-African gene flow into the Palenque are excluded from the analysis. E1b1a clade NRY are known to be strongly associated with the EBSP, which includes the Congo (Underhill et al. 2001; Cruciani et al. 2002; Wood et al. 2005; Berniell-Lee et al. 2009). The clade represented 87% of the Congolese dataset, 90% of Ghanaians, 76% of Malawians, 89% of Nigerians and 79% of the Mozambique collection. (Palenque 52%)

UEP+STR: characterising NRY at this level provides a fine level of resolution but suffers from the disadvantage that the number of shared chromosomes in cross group comparisons can be very low when groups are not closely related, either due to substantial time since a common origin or the absence of gene flow. As a consequence calculations of Fst can be driven by very few shared NRY, with comparative values of Fst not being a reliable indicator of which group, from a large list of candidates, is the most likely source. It is, however, an appropriate method for comparing Abajo with Arriba using ETPD, since the number of shared chromosomes is substantial (see Supplementary Table 2.S5 for NRY UEP+STR sharing across all groups).

Rst: in this method genetic distance among populations is assessed by analysing variation in STR repeat lengths and the measure has been shown to increase linearly over time groups have been separated. In the present study the short period since the foundation of Palenque de San Basilio makes it unlikely that mutation of STR is the important element in recognising the presence of a founding constituent within Palenque because it is the frequencies of shared lineages that is the best indicator of origin (see Thomas et al. 2000 and Veeramah et al. 2010) and Rst neither defines nor analyses lineage frequencies. Fst based on haplotype frequencies does utilise lineage frequencies and is, consequently, an appropriate method.

mtDNA VSO haplotypes: the VSO haplotype is suitable for analysis by Fst applied to haplotypes just as is the NRY haplotype. Over short time periods Fst between populations, based on the non-recombining VSO haplotype, is expected to increase in a similar way to Fst based on single-locus alleles, since variation in the HVR-1 sequence, due to nucleotide mutation, is not expected to greatly affect the frequencies of haplotypes, although there is likely to be some increase in the number of population specific haplotypes (Veeramah et al. 2010). Although it has been suggested that mtDNA haplogroups can be inferred from HVR-1 sequences (Richards et al. 1998; Simoni et al. 2000), this view has been strongly challenged (Torroni et al. 2000). Consequently, analysis was restricted to VSO haplotypes.

Nucleotide variation (π): calculating K2P between populations recognises both haplotypes and the quantum of variation among haplotypes, with increased difference being related to the time since separation. In doing so it takes into account, by allocations of different weightings, the frequency of transversion and transition changes. Calculations are therefore less likely to be affected by recent gene flow (resulting in low genetic distances) (Hebert et al. 2003).

Applying the above appraisal of methods of assessing similarity to the data answers to the questions posed in the aims of this study are as follows:

2.4.1.1. Are genetic data consistent with a Yombe speaking paternal founding group of the Palenque?

Fst based on UEPs between Palenque and 42 sub-Saharan groups in analysis that included haplogroups comprising the E1b1a clade was lowest between Palenque and Yombe. Because there was no significant difference between Yombe and Chewa at the all UEP and E1b1a characterisations a bootstrap simulation was performed in which the two Fst distances were compared following 100,000 re-samplings with Palenque –Yombe showing the smallest distance. Notably, in ETPD tests at both the all NRY clades and E1b1a levels Yombe were the only group that did not record a significant difference with the Palenque dataset (p=0.504). I also note that haplogroup E1b1a8a1a, while being absent in both the Palenque and all the Congo datasets (including Yombe), was present at more than 5% in the Chewa. Consequently, my answer to the question posed is in the affirmative. The two PCA plots in Figure 2.3 illustrate this showing a position of the Yombe closer to Palenque when the distances have been calculated by analysing all component haplogroups in just the E1b1a clade.

2.4.1.2. Is there any pattern in the distribution of maternal ancestry based on mtDNA HVR-1 haplotypes or K2P distances similar to that displayed by analysis of the NRY?

Due to the high level of genetic diversity observed at the HVR-1 locus in sub-Saharan African population groups, which is thought to be a consequence of patrilocality, and the hypervariable nature of this segment of mtDNA, genetic distances have less power to differentiate among population groups and comparisons made in this study clearly indicate a negative answer.

2.4.1.3. Can the two groups of residents of Palenque de San Basilio occupying Abajo and Arriba respectively be distinguished from each other and, if so, are the residents of Arriba more similar to the peoples of the Congo than are the residents of Abajo?

The two areas of the village can, on the basis of analysing NRY, be distinguished whether the analysis is undertaken using all NRY, just the E1b1a clade or UEP+STR (see Table 2.7). Quite contrary to this conclusion no significant difference was observed with respect to maternal inheritance.

When addressing the issue of similarity between the village sub-groups and African datasets we note that, analysing all NRY clades, the only non-significant Fst is between Abajo and Yombe, while for Arriba the least distance was between it and the Chewa. At the preferred E1b1a clade level in the cases of both Abajo and Arriba a

Congolese group was closest (Yombe and Bembe respectively). In both cases the Chewa were the second least distant group. Consequently, the genetic data do not provide support for the hypothesis that the inhabitants of the Arriba area have an ancestry more associated with the Yombe.

The genetic distance between the two districts was at its highest at the UEP level based only on E1b1a samples, with the distance between Abajo and Arriba greater than 85% of pairwise distances observed among sub-Saharan African groups. The proposition, based on sociological observation, that the original inhabitants of Arriba were first-generation slaves while the original inhabitants of Abajo were born into slavery in Colombia is not supported by the data. If the inhabitants of Abajo were born into slavery in Colombia it is to be expected, given the propensity of male slave owners to impregnate their female slaves, that they would have a far greater proportion of European NRY (haplogroup P*(xR1a) than does Arriba. This is not the case in Palenque (Abajo 15%, Arriba 19%, Fisher's exact test P = 0.639).

Analysis of mtDNA revealed no significant difference between Abajo and Arriba based on genetic distances (Fst and K2P, P > 0.712) and ETPD (P = 0.987). This lack of maternally mediated structuring is consistent with recent gene flow.

2.4.2 Additional observations

2.4.2.1 European and African contributions to the Palenque NRY gene pool

NRY haplogroup P*(xR1a) accounted for 18% of the individuals in Palenque, suggesting a substantial contribution by European males. Among the P*(xR1a) individuals twelve unique UEP+STR haplotypes were present, with haplotypes having as many as seven mutation steps distance from each other. This indicates that the contribution of this haplogroup to Palenque involved many different males. Furthermore, this haplogroup is most frequently observed in Spain, the colonial ruling power in 17th century Colombia (for the distribution of this haplogroup in Europe see Table 2.9).

Haplogroup E* (xE1b1a), observed at high frequencies in North and East Africa (see Table 2.9), is present at 20% in Palenque. It is very unlikely that haplogroups P*(xR1a) and E*(xE1b1a) were present among slaves born in the Congo as both haplogroups are present at very low frequencies in sub-Saharan Africa (especially in groups on the West Coast).

The presence of Spanish people, as colonial rulers, and the possible presence of slaves from North West Africa in Cartagena (taking into consideration that it was a multi-ethnic slave trade port in the 17th century) make it more likely that admixture occurred after the slaves had been brought to Colombia from sub-Saharan Africa. However, it should be noted that this does not exclude the possibility that these haplogroups were introduced into the Palenque community after its formation. Haplogroup R1a, for example, which represented only (3%, n = 4) in Palenque (all having the same UEP+STR haplotype), is at high frequency in East Europe and Central Asia (Underhill et al. 2000) and was most likely contributed by peoples of adjacent Colombian communities. Overall, the results indicate that a significant amount of admixture has occurred in the Palenque population.

Moreover, 3% of samples belonged to haplogroup Y*(xBR,A3b2), which is one of the oldest human NRY lineages and is thought to be the remnant of indigenous populations living in sub-Saharan Africa prior to the EBSP (Underhill et al. 2001). 52% of samples belong to the E1b1a clade (n = 78), which is almost absent from groups outside sub-Saharan Africa (see Table 2.9). The ‘Bantu modal haplotype’ (E1b1a-15-12-21-10-11-13) is not modal in the Palenque and the STR haplotype network does not have a star genealogy as observed in sub-Saharan African population groups. This pattern is consistent with a reasonably stable population and drift following formation. The described pattern of NRY types is therefore consistent both with a Yombe founding group and a substantial contribution of both African and non-African NRY. In the case of NRY haplogroups not usually observed in Africa, the respective NRY could, variously, have either been inherited from non-members of the village or when the NRY-carrying male joined the village.

Table 2.9. Frequency distribution of haplogroups P*(xR1a), E*(xE1b1a) and E1b1a in geographic locations across Europe, North Africa and East Africa.

Geographic location	P*(xR1a)	E*(xE1b1a)	E1b1a	N*	Reference
Spain					
Aragon	0.56	0.06	0.00	34	Adams et al. 2008
Andalusia (East)	0.72	0.04	0.00	95	
Andalusia (West)	0.56	0.15	0.00	73	
Asturias	0.50	0.15	0.00	20	
Basque country	0.88	0.01	0.00	116	
Castilla la Mancha	0.72	0.04	0.00	63	
Castile (North East)	0.77	0.09	0.00	31	
Castile (North West)	0.60	0.19	0.00	100	
Catalonia	0.81	0.03	0.00	80	
Extremadura	0.50	0.18	0.00	52	
Galicia	0.57	0.17	0.00	88	
Gascony	**0.97****	0.00	0.00	24	
Valencia	0.64	0.10	0.01	73	
Portugal					
North Portugal	0.59	0.15	0.00	60	Adams et al. 2008
South Portugal	0.47	0.17	0.01	78	
England					
Ashbourne	0.65	0.06	0.00	54	Weale et al. 2002
Southwell	0.64	0.07	0.00	70	
Fakenham	0.57	0.02	0.00	53	
North Walsham	0.58	0.04	0.00	26	
Wales					
Llangefni	0.89	0.04	0.00	80	Weale et al. 2002
Abergele	0.56	0.39	0.00	18	
Friesland	0.55	0.02	0.00	94	Weale et al. 2002
Algeria	0.06	0.57	0.00	54	Arredi et al. 2004
Morocco	0.03	**0.77**	**0.07**	146	Bosch et al. 2001
Tunisia	0.07	0.50	0.01	148	Arredi et al. 2004
Ethiopia	0.00	0.48	0.00	95	Moran et al. 2004

*N represents the number of individuals from each location. **Maximum frequency of each haplogroup is shown in bold type.

2.4.2.2 NRY distribution in sub-Saharan Africa

The E1b1a clade, as expected, is observed at high frequency in all 42 sub-Saharan African population groups. Although this NRY clade is modal in these groups, its frequency is variable among groups, ranging from 73% in South Africa to 96% in Nigeria and Cameroon.

Haplogroup Y*(xBR,A3b2) is found at its highest frequency in Khoisan (Underhill et al. 2000) and Cameroon Grassfields (Veeramah et al. 2008a) (44% and 23% respectively), but observed at very low frequencies with a patchy distribution in sub-Saharan Africa (Rosa et al. 2007 and Berniell-Lee et al. 2009). The frequencies of this haplogroup in all 42 groups characterised in this study (range of 0-5%) is in good agreement with published data and supports the idea that the expansion of Bantu-speaking peoples has dramatically influenced the pre-existing African NRY diversity.

Haplogroup E*(xE1b1a) was also present in most groups (range of 0-13%). The presence of haplogroup P*(xR1a) in sub-Saharan Africa at very low frequencies (range of 1-3%) may be the result of sporadic male introgression during the European colonisation of Africa.

2.4.2.3 Analysis of mtDNA variation in Palenque and sub-Saharan Africa

Gene diversity values, based on VSO haplotypes of the HVR-1 region, observed in sub-Saharan African groups are very similar to those found previously in West-Central African groups (Salas et al. 2002; Salas et al. 2004). Furthermore, the overall gene diversity in Congo was identical to that reported in the neighbouring country Angola (Plaza et al. 2004). Overall comparison of mtDNA gene diversities shows a significant level of similarity in areas affected by the expansion of Bantu-speaking peoples. Gene diversity in Palenque, compared with sub-Saharan African groups, is significantly lower ($P < 0.001$) and shows a significant level of drift in the maternal inheritance, with only five haplotypes covering 67% of the haplotypes. However, when diversity is estimated based on mean pairwise nucleotide differences (π, which takes into account the phylogenetic relationships of sequences), a different pattern is observed where seven sub-Saharan African groups have lower π values than

Palenque. This difference is consistent with mean pairwise nucleotide difference calculations being less sensitive to the effects of genetic drift; and values may even increase under such circumstances (Helgason et al. 2003).

2.4.2.4 Differentiation based on mtDNA

As explained earlier in the discussion of the merits of alternative methods of comparing multiple datasets the appropriate choice of method when assessing relative genetic distances between Palenque and multiple sub-Saharan groups using Fst calculated by analysing VSO haplotypes depends on an adequate number of common haplotypes either within a group or shared between groups. If haplotypes shared between groups are rare and have a wide geographic distribution the smallest distances may represent no more than the stochastic presence or absence of the rare haplotypes. The criterion of adequate sharing was arguably just satisfied in this study, with common haplotypes present in the Palenque dataset represented in 40.0% in the groups with which comparisons were made, and at between 7.3% to 16.6% of the samples. Genetic distances calculated using Fst and K2P are consistent both with multiple origins of the mtDNA observed in the Palenque and substantial gene flow between Abajo and Arriba.

2.4.2.5 Origins of Palenque

In seeking to identify the sub-Saharan group most similar in its paternal ancestry I took an approach similar to that taken by Di Giacomo et al. (2004), making use of detailed characterisation of chromosomes within the E1b1a haplogroup by recognising component haplogroups that comprise the clade. The five defined component-haplogroups were informative and permitted genetic distances between Palenque, Yombe and each of the other sub-Saharan groups to be discriminated. This approach to elucidating the origin-group of the Palenque is supported by the observation of the wide distribution and high frequency of the E1b1a clade in sub-Saharan Africa and its relative absence in males without a recent EBSP paternal ancestry.

Although both Fst and ETPD comparisons indicated that Yombe were the sub-Saharan group most similar to the Palenque there was an initially surprising observation that the Chewa (a group far to the east in Malawi) was only a little less similar. Furthermore, based on UEP-E1b1a, the Chewa-Palenque genetic distance was not significantly different from zero. The difference between Chewa and Yombe was also assessed by ETPD using UEP-E1b1a with no significant difference recorded. The modal haplogroup in Palenque (E1b1a8*) was also modal in both Yombe and Chewa, while all other groups (except Bantu speakers-Pretoria) had E1b1a7 as their modal haplogroup. Nevertheless, Yombe and Chewa did have differences in their NRY profiles in that the Chewa, unlike the Yombe, had NRY belonging to the E1b1a8a1a haplogroup (5.4% n = 5), a haplogroup that was not observed in the Palenque or any Congolese group. Notably, this component haplogroup is currently the most derived component haplogroup in the E1b1a clade, with an average frequency in south-east African groups in this study of 5.3%.

In *The History of the Chewa* (Ntara S.M. 1973) it is noted that, based on many Chewa traditions, the Chewa migrated from the Congo to Malawi approximately 500 years ago. This was of course only one or two hundred years before the Spanish began importing slaves from the Congo to Columbia. The most parsimonious explanation of the observed data is therefore that while the Yombe are the most closely related group to the Palenque the Chewa are in turn also closely related to both the Palenque and the Yombe, but did not contribute to the Palenque NRY gene pool since they had migrated out of the area before slaves were transported by the Spanish to Colombia. Presumably the Chewa acquired NRY belonging to the E1b1a8a1a haplogroup by introgression after they reached Malawi. The absence of NRY belonging to haplogroup E1b1a8a1a in both Yombe and Palenque further strengthens the conclusion that Yombe (a Congolese group) is a more likely origin group than are the Chewa (a Malawian group). However, it must be recognised that even though this study suggests a paternal Yombe origin for Palenque, this does not contradict the possibility that slaves from other parts of Congo joined the Palenque community either at its foundation or subsequently but that they were not in such number or authority so as to introduce the widespread use of their language.

The extent of genetic drift observed in Palenque and the presence of the high-frequency mtDNA haplotypes present in the Palenque in multiple sub-Saharan African groups at low frequencies made it impossible to suggest a candidate group as the most likely origin of Palenque. This is consistent with the oral history of the group which makes no mention of a founding group of females. It is also consistent with restricted absorption of women into the community and the retention of their female offspring, both scenarios being likely to produce the pattern of mtDNA haplotypes observed (i.e., a restricted number of types and high frequencies of those that are present). Notably, unlike the pattern observed for paternal descent the overwhelming majority of present day residents have a sub-Saharan maternal African ancestry.

2.4.3. Conclusion

This study, following similar methods of analysis to others that seek to trace origins of Diaspora (Thomas et al. 1998; Thomas et al. 2000; Thomas et al. 2002; Goncalves et al. 2007), has shown that sex-specific genetic systems have the potential to identify origins of peoples who have migrated thousands of miles and to trace their ancestry to a geographic location with a reasonable level of confidence, provided appropriate markers and methods are selected for the analysis.

The main findings of this study are:

1) The Palenque have substantial sub-Saharan African ancestry inherited along both the paternal and maternal lines.

2) Yombe, a Congolese group, is the most likely group from which the original settlers of Palenque, who have left present day descendents, came.

3) There is significant male-mediated genetic structuring distinguishing the two districts (Abajo and Arriba) of Palenque.

4) The analyses are in accordance with hypotheses based on linguistic and historical data, which suggested a) African Yombe speaking escaped slaves founded Palenque and b) the two areas of the village, Abajo and Arriba, have different paternal histories. They do not support the hypothesis that the original inhabitants of Arriba are more likely to have been born in the Congo than was the case for the original inhabitants of Abajo.

2.5. Supplementary Section for Chapter 2

Because of their large size, for Supplementary Tables 2.S1, 2.S2, 2.S3, 2.S4, 2.S5 and 2.S6 please see attached CD-ROM.

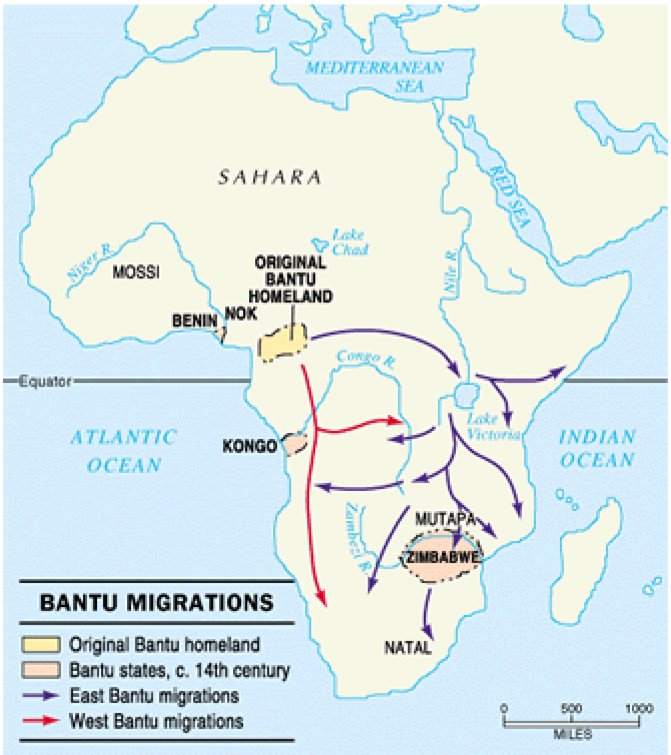
3. Genetic evidence pertaining to routes taken during the expansion of Bantu-speaking peoples (EBSP)

3.1. Introduction

3.1.1. Linguistic and archaeological analysis of EBSP

In many parts of the world, including Africa, the early development of agriculture has triggered significant population growth resulting in the expansion of early farming populations along with the expansion of language families (Bellwood 2001). This is mainly due to the many advantages of agricultural subsistence over foraging and thus agriculturists and their languages expanded at a revolutionary pace during the Holocene era (Diamond and Bellwood 2003). One unequivocal example of this phenomenon in Africa is the expansion of Bantu-speaking peoples (EBSP), which is thought to have started around 5000 years ago (Vansina 1995).

Figure 3.1. Suggested expansion routes taken during the expansion of the Bantu-speaking peoples, based on linguistic analysis



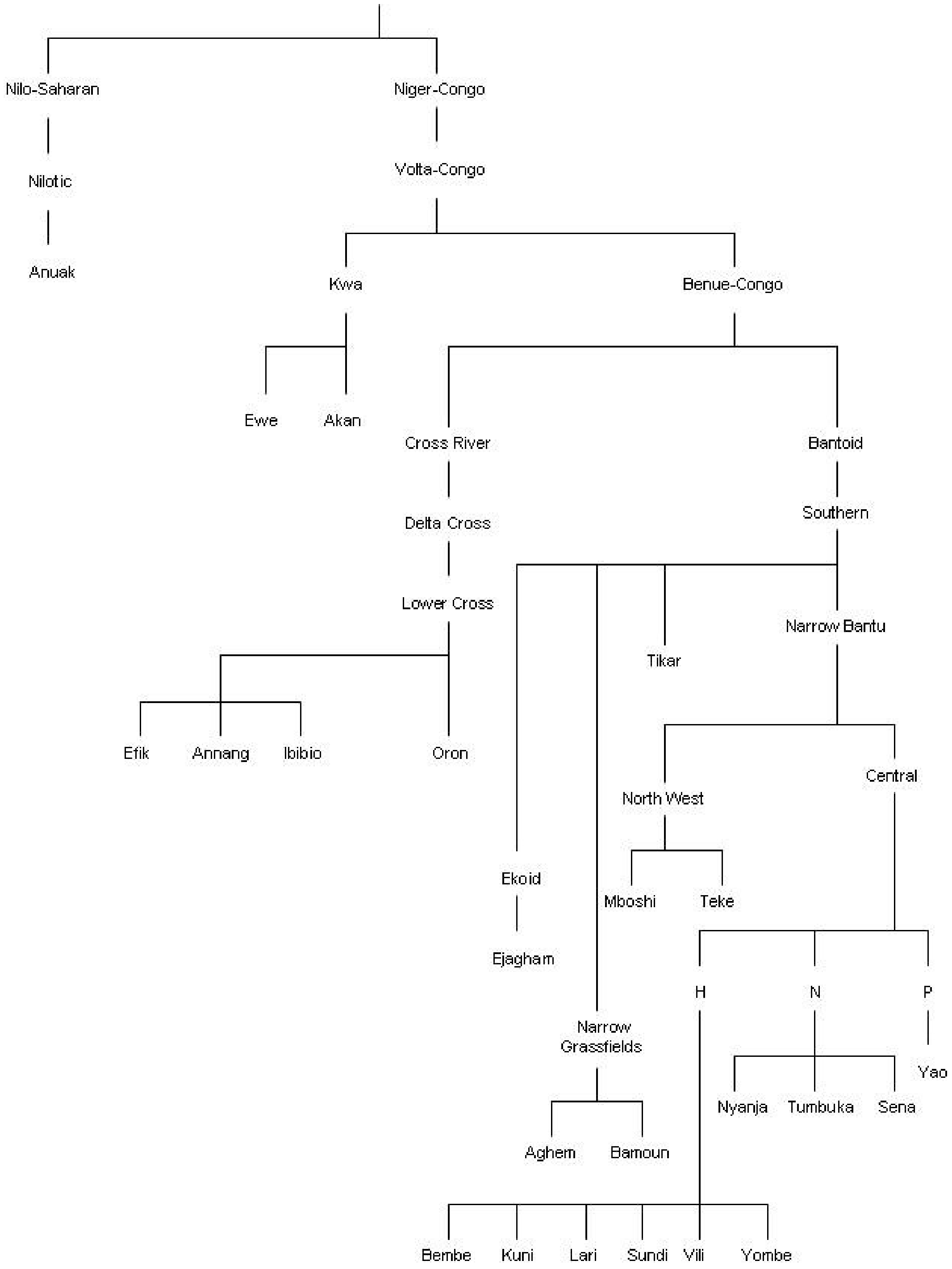
(http://www.public.asu.edu/~csteiner/).

Based on linguistic evidence, the region on the border between eastern Nigeria and Cameroon is thought to be the location of origin of proto-Bantu languages (Newman, 1995) (see Figure 3.1). Initially, farmers expanded east from this region and within 1,500 years reached West-Central Africa. After that the expansion took two directions with one wave moving along the south-western coast (West-Bantu route) and the other moving further east, forming the eastern Bantu core by 3,000 years before present (YBP) and moving along the south-eastern coast (East-Bantu route) (Pereira et al. 2001). The Bantu language subgroup that covers most of sub-equatorial Africa is, consequently, just one of the 177 subgroups of the Niger-Congo family (Diamond and Bellwood 2003) yet comprises a third of the languages that make up the largest phylum in the world, with more than 200 million speakers (Nurse 2006). This level of linguistic homogeneity among geographically distant populations across sub-Saharan Africa supports the notion that a rapid expansion of agriculturists has taken place. Although Bantu languages are similar, due to the separation of the two major waves around 3,500 YBP and the subsequent isolation of the various speakers, modern Bantu languages can be divided into West and East Bantu (Holden 2002; Jobling et al. 2004) (see Figure 3.2).

Within sub-Saharan Africa, areas that have not experienced significant levels of EBSP migration exhibit non-Bantu languages, which do not belong to the Niger-Congo phylum. These languages are mainly the Khoisanid languages in the southwest and a few Cushitic and Nilo-Saharan tongues in the northeast (Nurse 2006). It is also suggested that although the Bantu-speaking agriculturists may have replaced the local hunter-gatherers in their path they have also, in some places, co-existed and intermarried with the original inhabitants. These indigenous peoples are thought to be related to Khoisan and modern pygmies (Diamond and Bellwood 2003).

Based on archaeological evidence, the early expansion of proto-Bantu speakers was based on pre-Iron Age farming technology and did not involve smelting metals (Vansina 1995). The first sign of metallurgy south of the Sahara was found at Nok in Nigeria no earlier than 2,500 YBP (Cavalli-Sforza et al. 1994). Therefore, it is possible that, with the aid of the new technology, further expansions may have occurred after the first dispersal of farmers towards the south. Since the Bantu languages on the eastern route are shown to be more homogeneous compared with the western route (Excoffier et al. 1987), it is reasonable to suggest that further expansions have mainly occurred on the eastern route.

Figure 3.2. Broad relationships of the languages of groups referred to in this chapter based on Ethnologue (2009). Branch lengths are not informative.



3.1.2. Genetic analysis of EBSP

Early genetic studies of Bantu-speaking peoples were based on classical gene frequency data. Attempts were made to identify the genetic relationships among EBSP groups in the context of Africa as a whole (Excoffier et al. 1987 and Cavalli-Sforza et al. 1994). The major finding of these studies is that Genetic Distances (Fst) among all EBSP groups are much less than the average Fst among West African and Nilo-Saharan groups, indicating a considerable level of homogeneity among EBSP groups. More recently, based on over 1,300 autosomal markers, the genetic structure of African populations was analysed (Tishkoff et al. 2009). Results show that populations from areas affected by the EBSP exhibit a high level of genetic homogeneity, which is in good agreement with the findings of other recent studies (Verdu et al. 2009) and earlier studies mentioned above.

The EBSP impact on African demography has, over the past decade, mainly been studied by analysing paternal and maternal sex-specific genetic systems (non-recombining region of Y chromosome (NRY) and mitochondrial DNA (mtDNA) respectively). As both NRY and mtDNA genetic systems have smaller effective population sizes than autosomal markers, they are more prone to genetic drift (Roever et al. 2000; Destro-Bisol et al. 2004; Wood et al. 2005). Therefore, utilising NRY and mtDNA are more likely to reveal differences among groups than single linked autosomal markers.

The control region of the mtDNA sequence, due to its high mutation rate, has been extensively used in examining the impact of EBSP on the genetic landscape of sub-Saharan Africa (Pereira et al. 2001; Salas et al. 2002; Salas et al. 2004; Beleza et al. 2005). It has been postulated that some mtDNA haplogroups (e.g., L3b, L3e and L2a), based on their distribution in sub-Saharan Africa, are associated with the EBSP. The presence of haplogroup L1c at high frequency in some populations on the western route compared with the eastern route is thought to be the result of assimilation of local female hunter-gatherers (Salas et al. 2002). It has been suggested that since most agriculturist men marry local women and not vice versa (Destro-Bisol et al. 2004; Wood et al. 2005), the maternal genetic profile of EBSP groups is marked by high haplogroup diversity. Despite this level of diversity, great homogeneity exists among groups, which results in lack of genetic differentiation at the maternal level (Castri et al. 2009).

The increase in the rate of identification of slowly mutating NRY binary markers (i.e., Unique Event Polymorphisms (UEPs)) (Hammer et al. 1995, Underhill et al. 1997, Underhill et al. 2000) triggered many studies to investigate the paternally mediated genetic relationships of sub-Saharan African populations. Scozzari et al. (1999) and Underhill et al. (2001) found UEP (M2 and its analogues such as DYS271G) present at high frequencies specifically in sub-Saharan Africa and suggested this marker as a signature of EBSP. Since then, this marker (now defining the E1b1a haplogroup) has been widely typed in many groups across sub-Saharan Africa (Pereira et al. 2002, Cruciani et al. 2002, Beleza et al. 2005 and Coelho et al. 2009) and, without exception, all studies have shown that the majority of NRY types belong to this haplogroup. Although sampling in most NRY studies of sub-Saharan Africa has, in the past, been quite patchy in terms of geographic coverage and sample sizes, the distribution of this haplogroup is relatively well described in both major EBSP waves: southeast (Luis et al. 2004) and southwest (Coelho et al. 2009), and also in Senegal (Semino et al. 2002) and Cameroon (Cruciani et al. 2002, Luis et al. 2004) in West Africa.

Nevertheless, despite the considerable increase in the number of Y chromosome markers identified in recent years (Karafet et al. 2008) and the identification of further UEP within E1b1a (Sims et al. 2007), except E1b1a7 (M191; Underhill et al. 2001 and Cruciani et al. 2002), this very frequent haplogroup has not been further characterised in sub-Saharan African populations. As the EBSP shows a clearer genetic legacy in the paternally inherited genetic system compared with mtDNA (evident from high and similar frequencies of E1b1a) in sub-Saharan Africa (Berniell-Lee et al. 2009), it is possible that fine-scale E1b1a typing of Bantu-speaking communities throughout sub-Saharan Africa can shed light on routes taken during their expansion.

3.1.3. Aims

The main aims of this study are to investigate whether further characterisation of the NRY E1b1a haplogroup in 43 sub-Saharan African population groups with diverse linguistic affiliations (see Figure 3.2) would differentiate Bantu-speaking population groups and throw light on routes taken during the EBSP. Frequencies of haplogroups in different regions and variation in diversity of STR will be examined. Analysis of the distribution of diversity and estimates of the time to the most recent common ancestor (TMRCA) of UEPs defining clades of the E1b1a haplogroup have the potential to increase knowledge of the times when demographic events took place and the general location of those events. A secondary aim is to make an initial assessment of the potential for analysis of variation in mtDNA haplotypes to provide evidence of similar or different events to those suggested by NRY analysis.

3.2. Materials and Methods

3.2.1. Samples

Buccal swabs were collected from males over eighteen years old unrelated at the paternal grandfather level yet randomly selected from 43 locations across sub-Saharan Africa. These locations mainly cover West, Central-West, East, South-East and South Africa (see Table 3.1). All of the groups characterised in this study speak a Niger-Congo language except for the Anuak in south-west Ethiopia (representative of East Africa), who speak a Nilo-Saharan language.

Table 3.1. Details of population groups included in this study

Geographic location	Population Group (Code)	Ethnic Identity	Language	Latitude	Longitude	N
Cameroon	Bankim (BA)	Tikar	Tikar	6.083	11.483	33
	Foumban (FO)	Bamoun	Bamoun	5.729	10.902	117
	Wum (WU)	Aghem	Aghem	6.389	10.073	116
Congo	Bembe (BE)	Bembe	Bembe	-4.795	11.846	109
	Kuni (KU)	Kuni	Kuni	-4.795	11.846	68
	Lari (LA)	Lari	Lari	-4.259	15.285	62
	Mboshi (MB)	Mboshi	Mboshi	-4.259	15.285	91
	Sundi (SU)	Sundi	Sundi	-4.259	15.285	25
	Teke (TE)	Teke	Teke	-4.259	15.285	63

	Vili (VI)	Vili	Vili	-4.795	11.846	108
	Yombe (YO)	Yombe	Yombe	-4.432	12.108	65
Ethiopia	Anuak (AN)	Anuak	Anuak	7.953	34.412	108
Ghana	Asante (AS)	Asante	Akan	6.106	-1.878	94
	Ewe (EW)	Ewe	Ewe	6.6	0.467	88
	Fante (FA)	Fante	Akan	5.817	-2.817	60
Malawi	Chewa (CH)	Chewa	Nyanja	-13.607	33.918	92
	Tumbuka (TU)	Tumbuka	Tumbuka	-14.27	34.79	61
	Yao (YA)	Yao	Yao	-12.77	33.874	56
Mozambique	Sena (SE)	**Multiple**	Sena	-17.442	35.027	62
Nigeria	Abak (AB)	Annang	Annang	5.05	7.717	56
	Afaha Eket (AE)	Ibibio	Ibibio	4.717	7.867	48
	Afaha Okpo (AO)	Oron	Oron	4.833	8.233	50
	Afaha Ukwong (AU)	Oron	Oron	4.75	8.25	49
	Awa-Onna (AW)	Ibibio	Ibibio	4.69	7.815	28
	Calabar (CA)	Igbo	Igbo	4.95	8.317	99
	Ediene Ikono (ED)	Ibibio	Ibibio	4.783	7.883	48
	Efut Akpabuyo (EF)	Efik	Efik	4.908	8.442	48
	Efut Odukpani (EO)	Efik	Efik	5.167	7.983	49
	Ejagham-Akamkpa (EA)	Ekoi	Ejagham	5.35	8.35	47
	Ejagham-Calabar (EC)	Ekoi	Ejagham	4.95	8.317	83
	Eziagu Nenwe (EZ)	Igbo	Igbo	6.117	7.517	49
	Ikono (IK)	Annang	Annang	4.992	7.758	42
	Itam (IT)	Ibibio	Ibibio	5.042	7.842	50
	Nike Enugu (NE)	Igbo	Igbo	6.433	7.483	54
	Nnung Ndem-Onna (NN)	Ibibio	Ibibio	4.633	7.85	48
	Nsit (NS)	Ibibio	Ibibio	4.833	7.9	36
	Ntan Ibiono (NT)	Ibibio	Ibibio	5.233	7.933	50
	Obong Itam (OB)	Ibibio	Ibibio	5.133	7.967	50
	Oku-Itu (OI)	Ibibio	Ibibio	5.133	7.933	49
	Oku-Uyo (OU)	Ibibio	Ibibio	5.1	7.967	48
	Ukpom Ete (UE)	Ibibio	Ibibio	4.62	7.65	50
	Uwanse (UW)	Efik	Efik	4.95	8.317	50
South Africa	Bantu speakers-Pretoria (BN)	Bantu	Bantu	-25.746	28.187	98

All buccal swabs were collected anonymously with informed consent. Sociological data were also collected from most individuals including age, current residence, birthplace, self-declared cultural identity, first language, second language and (when available) religion of the individual, as well as similar information on the individual's father, mother, paternal grandfather and maternal grandmother. The samples were classified into groups primarily by cultural identity, first language spoken, then by place of collection. Where collections from a particular group were made in more than one location, locations are represented by averages of coordinates.

DNA from Congolese samples was extracted using the Gentra protein precipitation method (Gentra Systems, Minneapolis) (see Appendix A). Previously collected buccal swab DNA samples from ethnic groups across sub-Saharan Africa were extracted by standard phenol-chloroform method. The range and mean of sample sizes of the 43 groups are 28-118 and 63 respectively.

3.2.2. Y-chromosome typing

A combination of Unique Event Polymorphisms (UEP) and Short Tandem repeats (STRs) in the paternally inherited NRY was typed in eight Congolese groups (n = 597). The polymorphic markers are six STRs (DYS19, DYS388, DYS390, DYS391, DYS392 and DYS393) and four UEPs (M191, U175, U290 and U181) characterising the E1b1a haplogroup, which is modal in most population groups affected by the EBSP in sub-Saharan Africa (Underhill et al. 2001). The four UEPs were typed using a tetra primer ARMS PCR method (Ye et al. 2001), with minor modifications. The outer and two inner fragments were amplified in a 10-µl reaction volume containing 1 µl (~ 1 ng) of template DNA, 1.6 µl (50 µM) dNTPs, 9.3 nM TaqStartmonoclonal antibody (BD Biosciences Clontech, Oxford, UK), 1 µl of 10x Taq buffer, 0.13 units of Taq DNA polymerase (HT Biotech, Cambridge, UK) and outer and inner primers (see Table 3.2 for primer details). All samples (96-well plates) were then placed on a thermocycler under the following conditions: denaturation at 95°C for 5 min, followed by 35 cycles of denaturation (95°C) for 45 s, annealing (see Table 3.2 for annealing temperatures) for 45 s and elongation (72°C) for 45 s. The final step of the PCR program was a 7-min extension at 72°C before a 30-min hold at 4°C.

Where samples were not E1b1a (no derived alleles at the 4 UEP markers), a further six to eleven UEPs (TCGA UEP1 and UEP2 kits: sY81, SRY4064, YAP, SRY10831, M13, M9, SRY465, M20, Tat, 92R7 & M17) were typed (Thomas et al. 1999). NRY haplogroups were classified according to the nomenclature of the Y Chromosome Consortium (Karafet et al. 2008) and STR repeat sizes were assigned according to the nomenclature of Kayser et al. (1997).

Additionally, the four E1b1a-specific UEPs were also typed for previously characterised E1b1a samples in the TCGA database (unpublished data) from the other

35 sub-Saharan population groups (n = 1820) mentioned in Table 3.1. Although the battery of NRY markers typed in TCGA UEP kits gives a relatively crude resolution of NRY haplogroups, the typing of four UEP markers within E1b1a seems to considerably increase the resolution of NRY types associated with EBSP (Sims et al. 2007).

Table 3.2. Details of the primers used in the tetra-primer ARMS PCR reactions

UEP	Primer name	Primer sequence (5'-3')	Melting temperature (°C)	Fragment size (bp)	Primer concentration (µM)	PCR annealing temperature (°C)
M191 T>G	Outer forward primer	AATACCAGGCCGACATGGCAGCTA	64.4	399	0.15	61.5
	Outer reverse primer	CTACAAGCACGTACCACAGCGCCA	66.1		0.15	
	Inner forward primer	CATTTTTTTCTTTACAACTTGACCAG	56.9	133	0.75	
	Inner reverse primer	CACACCAAAATATCTCATATTTTCGTA	57.4	318	0.75	
U175 G>A	Outer forward primer	CCTTTAACACACTTCACAACATGG	59.3	274	0.3	53.0
	Outer reverse primer	GTGTCACTTTTCATTGTCTGG	55.9		0.3	
	Inner forward primer	CCACAGGTGCTAATGAAATCG	57.9	106	0.3	
	Inner reverse primer	ATGACCAGGAGAAGTCAAAAT	54.0	209	0.3	
U290 T>A	Outer forward primer	GCTATTGGAGAGCCTCGCTGTG	61.0	449	0.15	60.0
	Outer reverse primer	AGGAAGCAATTTTCCTACCTGCCA	59.5		0.15	
	Inner forward primer	GATAGGTGTGGGAATTGATGGCATT	58.3	193	0.75	
	Inner reverse primer	GATGGCCATCAGTCCCCAGT	60.4	300	0.75	
U181 C>T	Outer forward primer	GGTCTAGTGCACAGTGGTATCCA	57.1	389	0.15	62.0
	Outer reverse primer	AGAGCTCTCTCAAATCTGTGTTGG	57.5		0.15	
	Inner forward primer	AGTGTCTTTGTTTTGGCAAGAAC	61.8	123	0.75	
	Inner reverse primer	CTACCCTTGTATCAGAATACAGTTCTTA	61.7	316	0.75	

3.2.3. mtDNA typing

The mtDNA HVR-1 region of all Congolese groups and Chewa (Malawi) was sequenced as described by Thomas et al. (2002), but with the following modification: primers conL1-mod, conL2 and conH3 were replaced by conL849 (CTA TCT CCC TAA TTG AAA ACA AAA TA), conL884 (TGT CCT TGT AGT ATA A) and conHmt3 (CCA GAT GTC GGA TAC AGT TC) respectively for better sequencing results. For all samples, HVR-1 Variable Site Only (VSO) haplotypes were determined by comparing generated HVR-1 sequences of nucleotide range 16020-16400 with the Cambridge Reference Sequence (Anderson et al. 1981). It should be noted that Andrews et al. (1999) re-evaluated the mtDNA reference sequence by resequencing the entire mtDNA sequence and found no discrepancies in the HVR-1 region and therefore validated the use of the old sequence for the purposes of this study.

VSO haplotypes were defined by the type of occurring mutations (substitution, insertion or deletion) and their corresponding nucleotide positions. Haplogroups were classified based on the scheme of Salas et al. (2004) for samples of recent African ancestry. In order to extend the dataset, HVR-1 Variable Site Only (VSO) haplotypes were also determined for population groups previously typed in Nigeria, Ghana and Cameroon for the HVR-1 region (29 groups as shown in Table 3.1) (Veeramah et al. 2010). Unpublished HVR-1 data generated for the Anuak were also included as a representative of East Africa. As the nucleotide range of these samples for the HVR-1 region was 16023-16380, the HVR-1 coverage in Congolese samples was reduced to this range for comparisons with these groups.

The chromatogram of each sample was visually inspected across its whole length of sequence to check for high levels of background noise that would make it difficult to call bases. Subsequently, if no significant trace of noise was found, the ends of the sequence were then trimmed down to the standard nucleotide range (16020-16400) by visually inspecting the sequence. A contig is then formed for each set of 96 samples (i.e., one sequencing run) and each mutation site was examined visually to determine whether it is a substitution, insertion or deletion. All samples with any ambiguous sites were resequenced again.

3.2.4. Statistical analysis

Gene diversity, *h,* and its standard error were estimated from unbiased formulae of Nei (1987). Mean number of pairwise nucleotide differences were calculated for mtDNA HVR-1 sequences in each group. Population genetic structure was estimated using Hierarchical Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992), which takes into account the evolutionary relationship between pairs of haplotypes and generates a statistic called Fixation Index (F_{ST}), when a simple structure of populations within a single group is defined. However, when a hierarchical structure is defined three fixation indices, F_{ST} (among all groups), F_{SC} (among groups within a geographic region) and F_{CT} (among geographic regions), are generated. Patterns of genetic differentiation were quantified using the following genetic distance measures, estimated from AMOVA based φ-st values (Excoffier et al. 1992 and Michalakis & Excoffier 1996): a) F_{ST} (Reynolds, Weir & Cockerham 1983) (based on UEP haplogroups and mtDNA HVR-1 VSO haplotypes) and b) the Kimura's two-parameter model with gamma distribution of value 0.47 (Kimura 1980) (based on mtDNA HVR-1 sequences). Significance of Fixation Indices and genetic distances were assessed by permutation test where a null distribution is formed by calculating all possible values under rearrangements of haplotypes in the observed samples. The permutation test was repeated over 1,000 times to give rise to the null distribution. All the above was performed using Arlequin software version 3.0 (Excoffier et al. 2005).

Principal Component Analysis (PCA) biplot was performed using the 'R' environment of statistical computing (www.R-project.org), by implementing the 'princomp' function based on E1b1a component haplogroup frequencies and visualised using the 'biplot' function. The plot was used to visualise relationships among groups and identify clusterings with respect to the categorical variable (i.e., E1b1a component haplogroups). Average Squared Distances (ASD) were calculated in MS Excel and corresponding 95% confidence intervals were calculated as in Thomas et al. (1998) using 'R'. Time to Most Recent Common Ancestor (TMRCA) was estimated using an average NRY STR mutation rate of 0.002 (Veeramah et al. 2010) and generation time of 25 years.

3.3. Results and Discussion

3.3.1. NRY distribution in sub-Saharan Africa

Based on TCGA UEP1 and UEP2 kits, 13 NRY haplogroups can be defined. In the present dataset (n = 2,757) eight haplogroups were present, with E1b1a being modal across all groups (see Table 3.3). This haplogroup was in the range of 38.9-95.7%, with mean of 85.2 % and variance of 0.9%. The pattern of NRY observed in these 43 groups is in accordance with previously published data on sub-Saharan Africa

(Underhill et al. 2001; Cruciani et al. 2002; Wood et al. 2005; Luis et al. 2004; Berniell-Lee et al. 2009). One other major haplogroup observed in most groups is E*(xE1b1a), which is found at high frequencies north of the Sahara (Bosch et al. 2001; Arredi et al. 2004), consistent with gene flow across the Sahara into sub-Equitorial Africa.

Table 3.3. Haplogroup frequencies in 43 sub-Saharan African population groups

NRY UEP Haplogroup (according to the nomenclature of the Y-chromosome consortium (2008))	**P*(xR1a)**	**BT*(xDE,KT)**	**E*(xE1b1a)**	**K*(xL,N1c,O2b,P)**	**Y*(xB,A3b2)**	**DE*(xE)**	**A3b2**	**E1b1a**
Bembe (BE)	0.018		0.110				0.018	**0.853**[a]
Kuni (KU)		0.044	0.074		0.015			**0.868**
Lari (LA)		0.032	0.113					**0.855**
Mboshi (MB)	0.011	0.044			0.033			**0.912**
Sundi (SU)		0.040	0.080					**0.880**
Teke (TE)	0.032	0.063	0.079		0.016			**0.810**
Vili (VI)	0.009	0.028	0.019					**0.944**
Yombe (YO)	0.015	0.046	0.092		0.031			**0.815**
Congo Total	0.012	0.034	0.066		0.012		0.003	**0.873**
Foumban (FO)		0.085	0.043		0.009		0.009	**0.855**
Bankim (BA)	0.030	0.061	0.061					**0.848**
Wum (WU)		0.043						**0.957**
Cameroon Total	0.004	0.064	0.026		0.004		0.004	**0.898**
Asante (AS)			0.043		0.053			**0.904**
Fante (FA)	0.033		0.017		0.017			**0.933**
Ewe (EW)	0.023	0.023	0.045					**0.909**
Ghana Total	0.017	0.008	0.037		0.025			**0.913**
Chewa (CH)		0.076	0.130	0.011			0.022	**0.761**
Tumbuka (TU)		0.082	0.098	0.016			0.016	**0.787**
Yao (YA)		0.196	0.071					**0.732**
Malawi Total		0.110	0.105	0.010			0.014	**0.761**
Abak (AB)		0.054	0.036					**0.911**
Afaha Eket (AE)		0.125	0.042					**0.833**
Afaha Okpo (AO)		0.020	0.060			0.020		**0.900**
Afaha Ukwong (AU)		0.102	0.020		0.020			**0.857**
Awa-Onna (AW)		0.179	0.036					**0.786**
Calabar (CA)		0.071	0.010		0.010	0.010		**0.899**
Ediene Ikono (ED)		0.125	0.063					**0.813**
Efut Akpabuyo (EF)	0.021	0.188	0.021					**0.771**
Efut Odukpani (EO)		0.061	0.041					**0.898**
Ejagham-Akamkpa (EA)	0.021	0.085	0.043					**0.851**
Ejagham-Calabar (EC)	0.024	0.060	0.024					**0.892**
Eziagu Nenwe (EZ)		0.041						**0.959**
Ikono (IK)		0.167	0.024					**0.810**
Itam (IT)		0.040			0.020			**0.940**
Nike Enugu (NE)		0.056	0.019			0.019		**0.907**
Nnung Ndem-Onna (NN)		0.021	0.021			0.021		**0.938**
Nsit (NS)		0.056	0.028					**0.917**
Ntan Ibiono (NT)		0.080	0.020					**0.900**
Obong Itam (OB)	0.020	0.100	0.040	0.020		0.020		**0.800**
Oku-Itu (OI)		0.122						**0.878**

Oku-Uyo (OU)		0.042	0.063					**0.896**
Ukpom Ete (UE)		0.080	0.040			0.020		**0.860**
Uwanse (UW)		0.100			0.020			**0.880**
Nigeria Total	0.004	0.082	0.027	0.001	0.003	0.005		**0.877**
Bantu speakers-Pretoria (BN)		0.153	0.051	0.010	0.031			**0.755**
Sena (SE)		0.081	0.129					**0.790**
Anuak (AN)		0.250	0.167				0.194	**0.389**
Grand Total	0.006	0.075	0.051	0.001	0.008	0.002	0.010	**0.847**

[a] Modal haplogroups are shown in bold type

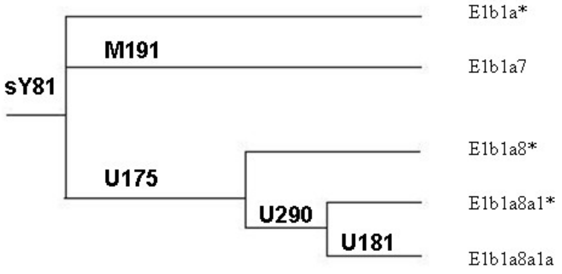
Furthermore, 7.5% of all samples belong to the broad haplogroup BT*(xDE, KT), which reaches high frequency on the eastern coast of Africa from Ethiopia to South Africa (see Table 3.3). This haplogroup is present at higher frequencies in Europe (Scozzari et al. 2001 and Weale et al. 2002) and Asia (Weale et al. 2001) and may have entered Africa through back-migration. NRY haplogroup P*(xR1a) is most frequent outside Africa and reaches high frequency in West Europe (Weale et al. 2002; Adams et al. 2008). Therefore, the sporadic presence of this haplogroup in groups located in West-Central Africa (see Table 3.3) is also thought to be the result of back-migration into Africa (Cruciani et al. 2002). As the haplogroups mentioned above (except E1b1a) are mostly broad paragroups, no inferences were made based on their frequency distributions since it could result in conclusions which could be misleading (Weale et al. 2003).

In-depth analysis of appropriate NRY haplogroups has been shown to be informative in understanding the peopling of a continent (Di Giacomo et al. 2004). Since E1b1a is modal in all groups investigated in this study (see Table 3.3) and also extensively reported to be associated with the EBSP (Underhill et al. 2001; Cruciani et al. 2002; Pereira et al. 2002; Berniell-Lee et al. 2009), this study was directed to the analysis of samples belonging to the E1b1a haplogroup (E1b1a dataset; n =2,336, 84.7% of original samples), which were genotyped to assign chromosomes to the component haplogroups of E1b1a.

3.3.2. Further characterisation of the E1b1a haplogroup

A further four UEP within E1b1a (see Figure 3.3) were typed in the E1b1a dataset samples in a hierarchical manner (see Table 3.4). These UEP cover most of the internal diversity within the E1b1a haplogroup observed in the HapMap Build 36 Yoruba (YRI) dataset (www.hapmap.org) and a population of African Americans (Sims et al. 2007).

Figure 3.3. Phylogenetic tree of E1b1a component haplogroups characterised in this study.



Nei's Gene diversity, based on E1b1a component haplogroups, for the whole set was 0.678 ± 0.007. Except for Anuak, where all samples are E1b1a7 (h = 0), gene diversity ranged from 0.379-0.753 in individual groups (mean of 0.645 and variance of 0.007), with Foumban and Fante having the lowest and highest values respectively. This indicates that, although population groups in sub-Saharan Africa all have a high frequency of E1b1a, there is considerable internal variation within this haplogroup (see Figure 3.4), which potentially allows the groups to be differentiated analysing only the E1b1a dataset.

The modal STR haplotype (comprised of six STR loci) in 36 groups was the EBSP modal STR haplotype (15-12-21-10-11-13) (Thomas et al. 2000; Pereira et al. 2002), with a frequency range of 0.143-0.396, a mean of 0.233 and variance of 0.004. In the other seven groups (Anuak, Bankim, Foumban, Sena, Sundi, Vili and Wum) the EBSP modal haplotype had a frequency range of 0.024-0.147 and the modal haplotype a range of 0.143-0.667, one to four mutation steps away from the EBSP modal STR haplotype.

Figure 3.4. Visual representation of E1b1a component-haplogroup distributions in sub-Saharan African populations.

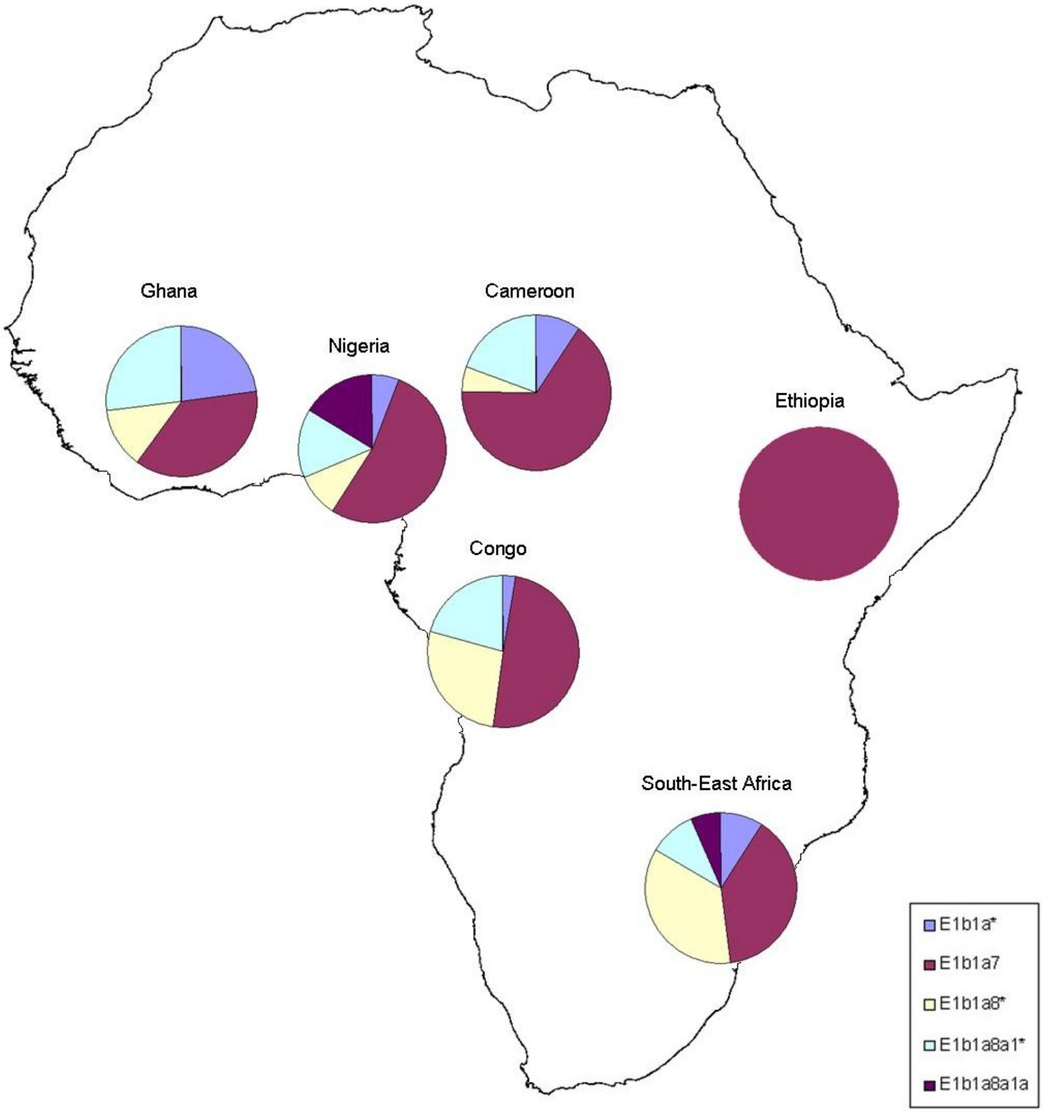


Table 3.4. Frequency distribution of five E1b1a component haplogroups defined in this study.

NRY UEP Haplogroup (according to the nomenclature of the Y-chromosome consortium (2008))	**E1b1a***	**E1b1a7**	**E1b1a8***	**E1b1a8a1***	**E1b1a8a1a**
Bembe (BE)	0.043	**0.376**[a]	0.247	0.333	
Kuni (KU)	0.017	**0.593**	0.186	0.203	
Lari (LA)		**0.566**	0.208	0.226	
Mboshi (MB)		**0.482**	0.313	0.205	
Sundi (SU)	0.045	**0.682**	0.227	0.045	
Teke (TE)		**0.529**	0.275	0.196	
Vili (VI)	0.059	**0.529**	0.265	0.147	
Yombe (YO)	0.038	0.340	**0.453**	0.170	
Congo Total	0.027	**0.492**	0.273	0.207	
Bankim (BA)	**0.429**	0.214	0.179	0.179	
Foumban (FO)	0.080	**0.780**	0.080	0.060	
Wum (WU)	0.027	**0.649**		0.324	
Cameroon Total	0.096	**0.653**	0.054	0.197	
Asante (AS)	0.271	**0.341**	0.118	0.271	
Ewe (EW)	0.125	**0.463**	0.113	0.300	
Fante (FA)	**0.321**	0.268	0.179	0.232	
Ghana Total	0.231	**0.367**	0.131	0.271	
Chewa (CH)	0.057	0.314	**0.386**	0.171	0.071
Tumbuka (TU)	0.042	**0.521**	0.292	0.104	0.042
Yao (YA)	0.098	**0.439**	0.341	0.073	0.049
Malawi Total	0.063	**0.409**	0.346	0.126	0.057
Abak (AB)	0.059	**0.392**	0.118	0.176	0.255
Afaha Eket (AE)	0.125	**0.425**	0.050	0.150	0.250
Afaha Okpo (AO)	0.022	**0.511**	0.067	0.089	0.311
Afaha Ukwong (AU)	0.048	**0.476**	0.119	0.167	0.190
Awa-Onna (AW)		**0.636**		0.227	0.136
Calabar (CA)	0.067	**0.528**	0.135	0.112	0.157
Ediene Ikono (ED)	0.103	**0.590**	0.051	0.103	0.154
Efut Akpabuyo (EF)	0.162	**0.568**	0.054	0.054	0.162
Efut Odukpani (EO)	0.023	**0.500**	0.159	0.091	0.227
Ejagham-Akamkpa (EA)	0.025	**0.500**	0.100	0.325	0.050
Ejagham-Calabar (EC)	0.027	**0.581**[a]	0.054	0.216	0.122
Eziagu Nenwe (EZ)	0.298	**0.489**	0.021	0.106	0.085
Ikono (IK)		**0.529**	0.206	0.176	0.088
Itam (IT)	0.085	**0.553**	0.064	0.128	0.170
Nike Enugu (NE)	0.020	**0.755**		0.082	0.143
Nnung Ndem-Onna (NN)	0.044	**0.600**	0.067	0.267	0.022
Nsit (NS)	0.030	**0.394**	0.121	0.273	0.182
Ntan Ibiono (NT)	0.044	**0.422**	0.178	0.156	0.200
Obong Itam (OB)		**0.600**	0.075	0.175	0.150
Oku-Itu (OI)		**0.581**	0.116	0.116	0.186
Oku-Uyo (OU)	0.047	**0.442**	0.093	0.233	0.186
Ukpom Ete (UE)	0.070	**0.558**	0.163	0.116	0.093
Uwanse (UW)	0.068	**0.523**	0.068	0.068	0.273

Nigeria Total	0.061	**0.529**	0.092	0.153	0.165
Bantu speakers-Pretoria (BN)	0.081	0.338	**0.378**	0.095	0.108
Sena (SE)	0.204	**0.388**	0.347	0.041	0.020
Anuak (AN)		**1.000**			
Grand Total	0.076	**0.509**	0.162	0.172	0.081

[a] Modal sub-haplogroups within E1b1a are shown in bold type
Note: Where a haplogroup is absent (frequency of zero), the corresponding cell is shaded in grey.

Notably, when haplotypes were defined by both UEP and STR markers combined the EBSP modal haplotype was observed almost exclusively within the E1b1a8 clade (i.e., the sum of E1b1a8*, E1b1a8a1* and E1b1a8a1a) rather than E1b1a7 (Fisher's exact test p-value << 0.00001). This finding is consistent with the scarcity of the EBSP modal haplotype in the seven groups mentioned above since they have a much smaller E1b1a8 proportion than any other group in this study.

3.3.3. Can geographic regions in sub-Saharan Africa and individual population groups within them be differentiated based on E1b1a UEP markers?

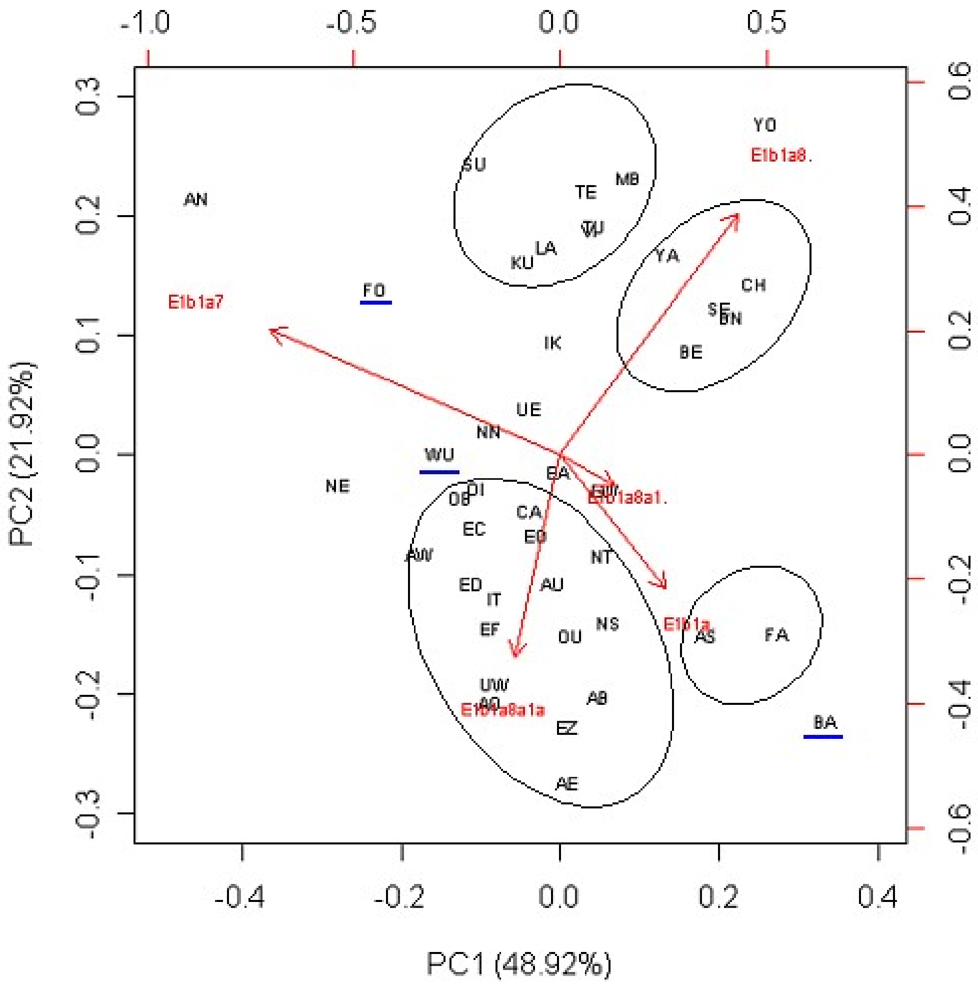
The AMOVA based global fixation index (F_{ST}) for the E1b1a dataset (43 groups) was 0.0694 and highly significant (P<0.0001), displaying substantial genetic heterogeneity among population groups.

Principal component analysis biplot was performed to identify population structuring in sub-Saharan Africa, based on all individual population groups (see Figure 3.5). On the basis of the first two dimensions, which capture 72% of variation, populations were grouped into four major clusters corresponding to Nigerian, Ghanaian, Congolese and South-East African groups.

In constrast, Cameroonian groups did not show any clustering. The main determinants (explanatory variables) of these clusterings were the E1b1a component haplogroups, except E1b1a8a1*. Among the four clusters, Congolese and South-East African groups showed the highest level of similarity (based on E1b1a8* frequency).

Groups clustered based on E1b1a* were mainly Ghanaian, whereas groups clustered based on E1b1a8a1a were mainly Nigerian. Furthermore, based on the defined E1b1a component haplogroups, a considerable degree of inter-population diversity is also observed within clusters.

Figure 3.5. PCA biplot of 43 sub-Saharan African population groups based on E1b1a component haplogroups.



Note: Percentages in parentheses are the amount of variation explained by each principal component. Cameroonian groups are indicated by a blue underline.

Groups were then analysed in separate groupings, based on the geographic clusters observed, and AMOVA based fixation indices (F_{ST}) were estimated (see Table 3.5). The low figures of F_{ST} observed in four geographic regions are consistent with either a recent common origin or high inter-group gene flow. The large significant Fst value in Cameroon is consistent with the observation in the PCA biplot.

Furthermore, the level of genetic structure was also examined among the geographic groupings in a pairwise manner (see Table 3.6). All pairwise distances were highly significant (P < 0.0001) and population structuring was evident. In particular, Anuak had the highest degree of differentiation from all other groupings (Fst > 0.163).

Table 3.5. AMOVA fixation indices in five geographic regions in sub-Saharan Africa

Geographic location	AMOVA F_{ST}	P-value
Ghana (n = 3)	0.016	**0.038**
Nigeria (n = 23)	0.014	**0.002**
Cameroon (n = 3)	0.177	**<0.0001**
Congo (n = 8)	0.019	**0.003**
South East Africa* (n = 5)	0.004	0.234

Significant P-values at the 5% level are shown in bold type. *This region includes the following groups: CH, TU, YA, SE and BN.

It should be noted from other studies that the Anuak are not to be considered representative of Ethiopia (Browning 2009). The main body of the Anuak live in southern Sudan. They were included in this study because they are an eastern Africa Nilo-Saharan speaking people with a considerable proportion of E1b1a chromosomes.

Table 3.6. AMOVA-based fixation indices (F_{ST}) among geographic locations and P-values.

	CAM	CON	ETH	GHA	NIG	SE
CAM		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
CON	0.060		<0.0001	<0.0001	<0.0001	<0.0001
ETH	0.163	0.266		<0.0001	<0.0001	<0.0001
GHA	0.082	0.056	0.359		<0.0001	<0.0001
NIG	0.034	0.048	0.193	0.068		<0.0001
SE	0.127	0.025	0.346	0.065	0.072	

Lower left triangle reports AMOVA-based F_{CT} values and the upper right triangle reports the corresponding P-values. [a] Abbreviations for the geographic regions are : CAM: Cameroon, CON: Congo, GHA: Ghana, NIG: Nigeria, SE: South East Africa and ETH: Ethiopia.

Overall, the F_{ST} results suggest that, based on UEP variation within E1b1a, there is an affinity between the West African and South East African groupings, which cover an area that includes both the proposed place of origin and of expansion of the Bantu-speaking peoples.

Table 3.7 records the presence or otherwise of the component haplogroups of E1b1a in the different regions of sub-Saharan Africa. While the Anuak of Ethiopia only displayed haplogroup E1b1a7 all the haplogroups were represented in Nigeria and the three south east African sample sets (Malawi, Mozambique and South Africa). Notably, Ghana, Cameroon and Congo all lacked haplogroup E1b1a8a1a.

3.3.4 Estimating TMRCA of the clade-defining UEPs

The TMRCA for each clade defining UEP was estimated based on ASD since TMRCA and ASD have a linear relationship with mutation rate as the constant (ASD = μt) (Thomas et al. 1998). They are recorded in Table 3.7 and are completely consistent with the genealogy in Figure 3.3, with the most recent dates associated with the finer branches. The TMRCA for the UEP defining the entire E1b1a clade was estimated at between 6,200-6,600 YBP which is more than a millennium before the proposed start of the EBSP (Berniell-Lee et al. 2009).

The TMRCA for E1b1a7 was estimated at between 4,800 – 5,300YBP, in good agreement with the linguistically-based time of the commencement of the EBSP. This component haplogroup was modal in 38 groups across sub-Saharan Africa, with the highest frequencies observed in Foumban (Cameroon), Nike Enugu (Nigeria) and Anuak (Ethiopia). Overall, the data are consistent with the M191 mutation initially appearing around 5,000YBP in present day Nigeria, since the highest STR gene diversity was observed in this region (h = 0.931).

TMRCA for the E1b1a8 clade was estimated at 1,900 – 2,200YBP, which is very similar to the earlier estimate by Veeramah et al. (2010). Since this clade is widespread across many areas of sub-Saharan Africa (especially west of Nigeria in Ghana) and the highest STR gene diversity was found in Nigeria (h = 0.885), it is likely that the U175 mutation arose in this area about 2,000 YBP and expanded both west and south. The wide distribution of E1b1a8 suggests that the expansion of the U175 mutation took place rapidly, which may be a consequence of the development of metallurgy in the region around 2,500 YBP (Cavalli-Sforza et al. 1994), initiating a further expansion of people, this time with some carrying the U175 mutation.

With respect to people belonging to sub-haplogroup E1b1a8* (only carrying the U175 mutation and not U290, U181 or both), the data indicate a substantial presence in all regions (>40% excluding Cameroon, 25.1% Cameroon). The TMRCA of the U290 mutation (defining the E1b1a8a1 clade) was dated to between 1,400 – 1,700 YBP. This mutation is also widespread, falling below 20% only in the Sena dataset representing Mozambique.

The TMRCA for U181, which is diagnostic of the most derived element of the E1b1a clade, component haplogroup E1b1a8a1a, characterised in this study, was estimated at 1,100 – 1,600 YBP. Interestingly, unlike the other E1b1a component haplogroups, E1b1a8a1a did not have a widespread presence in sub-Saharan Africa and was only found in groups located in Nigeria and south-east Africa. Based on STR gene diversity within E1b1a8a1a (Nigeria (h = 0.626) and South East Africa (h = 0.399)), it is more likely that this haplogroup originated in Nigeria than in South East Africa. Therefore, the data support an interpretation that men with NRY of the E1b1a8a1a haplogroup initially migrated east and then southwards on the eastern coast. It is unlikely that they would have taken a western route without leaving representatives of this haplogroup in the sample sets from Congo. The haplogroup appears to have a recent origin and NRY belonging to it have expanded at a very high rate of about 90km per generation (i.e., ~ 5000km in 55 generations) along the eastern route.

Finally, the results also suggest that significant NRY flow from the eastern coast to the west coast is unlikely during the past 1,000 years or so since if it had occurred, the presence of E1b1a8a1 NRY on the western coast would be expected. The association of the eastern route with haplogroup E1b1a8a1a could be further examined by typing samples from Kenya and Tanzania in East Africa, where the E1b1a haplogroup has been reported as modal (Luis et al. 2004).

Table 3.7. Details of E1b1a UEP internal diversity, estimated TMRCA dates and distribution of E1b1a component haplogroups in sub-Saharan Africa along with STR gene diversity within component haplogroups.

Clade defining marker	Gene Diversity (h) (SD)	ASD (95% CI)	TMRCA (YBP)	Haplogroup(s)	Geographic presence and gene diversity					
					West Africa			West Central Africa	East Africa	South East Africa
					Ghana	Nigeria	Cameroon	Congo	Ethiopia	Total
sY81	0.925 (0.003)	0.510 (0.494-0.527)	6,175-6,588	E1b1a* [a]	YES	YES	YES	YES	NO	YES
				E1b1a7	YES	YES	YES	YES	YES	YES
				E1b1a8*	YES	YES	YES	YES	NO	YES
				E1b1a8a1*	YES	YES	YES	YES	NO	YES
				E1b1a8a1a	NO	YES	NO	NO	NO	YES
M191	0.920 (0.003)	0.404 (0.384-0.424)	4,800-5,300	E1b1a7	YES (h=0.879)	YES (h=0.931)	YES (h=0.888)	YES (h=0.875)	YES (h = 0.516)	YES (h=0.911)
U175	0.775 (0.014)	0.161 (0.149-0.173)	1,863-2,163	E1b1a8*	YES (h=0.874)	YES (h=0.885)	YES (h=0.846)	YES (h=0.799)	NO	YES (h=0.784)
				E1b1a8a1*	YES	YES	YES	YES	NO	YES
				E1b1a8a1a	NO	YES	NO	NO	NO	YES
U290	0.679 (0.021)	0.125 (0.113-0.138)	1,413-1,725	E1b1a8a1*	YES (h=0.781)	YES (h=0.606)	YES (h=0.802)	YES (h=0.584)	NO	YES (h=0.763)
				E1b1a8a1a	NO	YES	NO	NO	NO	YES
U181	0.608 (0.040)	0.109 (0.088-0.131)	1,100-1,638	E1b1a8a1a	NO	YES (h=0.626)	NO	NO	NO	YES (h=0.399)

[a] Since E1b1a* is a paragroup and making inferences based on its distribution may be misleading, STR gene diversity values were not calculated.

3.3.5. The distribution of mtDNA haplotypes across sub-Saharan Africa

The HVR-1 region of mtDNA was sequenced in a total of 683 samples from eight Congolese groups and Chewa in Malawi. HVR-1 sequencing data for 29 groups representing Cameroon (n =3), Ghana (n =3) and Nigeria (n = 23) were taken from Veeramah et al. (2010) and HVR-1 data for Anuak (Ethiopia) was taken from the TCGA database (unpublished data). Although assignments to mtDNA haplogroups have been made based only on the HVR-1 sequence (Richards et al. 1998), this was not done in this study because there are reports that such assignments are not robust (Torroni et al. 2000). However, in another study of over 300 samples from Ethiopian groups in which both the HVR-1 sequence and 23 haplotype-defining markers from the coding region were typed only one case of homoplasy of the HVR-1 sequence across haplogroups was observed and that involved two phylogenetically closely related clades (unpublished PhD thesis in preparation, Plaster 2010). The analysis in this section is based on HVR-1 sequences without imputation of haplogroups. In the entire dataset, a total of 792 VSO haplotypes were identified, of which 477 were singletons (60.2%) while the modal haplotype represented 0.03 of the total. Nei's gene diversity for the whole dataset was very high (h = 0.994 ± 0.0003). Nei's gene diversity in the various datasets ranged from 0.968-0.997, with a mean of 0.987 and variance of 0.00005. The range of gene diversity observed in these groups is very similar to those observed previously in sub-Saharan African groups (Salas et al. 2002; 2004). Mean number of pairwise nucleotide differences (π) in the whole dataset was 8.85 ± 4.08 and ranged from 6.69 - 9.98 in individual groups, with a mean of 8.68 and variance of 0.55.

Of the 315 haplotypes that were not singletons only 45 were represented by 10 or more samples (see Table 3.8). Of these only one, a singleton, was observed in the South East Africa dataset but not in the Congo and all but two haplotypes were observed in the West African dataset (Ghana, Nigeria and Cameroon). Two of the ten or more mtDNA haplotypes were observed only in the Congo dataset (see Table 3.8). There were two non-singleton haplotypes (n=7, ~10%) observed in the South East African dataset that were not seen in any of the other datasets.

AMOVA-based fixation indices (F_{ST}, F_{SC} and F_{CT}) were estimated, based on the six geographic regions defined earlier, using VSO haplotypes. The fixation index F_{ST} (among population groups) was very low (0.0088), yet highly significant ($p < 0.0001$). The values of fixation indices F_{CT} (0.0054, $p < 0.0001$) and F_{SC} (0.0034, $p < 0.0001$) indicate that more genetic variation is present among the defined geographic regions than among groups in a single region. When the fixation indices were computed based on Kimura's 2P (K2P) distance (using HVR-1 sequences), the variation among groups (F_{ST}) increased to 0.027 ($p < 0.0001$). According to the results, most of the variation was found to be among the geographic regions ($F_{CT} = 0.023$, $p < 0.0001$) rather than within each region ($F_{SC} = 0.004$, $p < 0.0001$). The higher level of differentiation observed is probably due to the incorporation of the phylogenetic relationship of sequences (with the correction of nucleotide substitution rates for transitions and transversions) by K2P, which VSO haplotypes do not take into account.

Table 3.8. VSO haplotypes of ten or more total counts in different regions of sub-Saharan Africa.

VSO haplotype	Total West Africa	Total Congo	East Africa	South East Africa	Grand total
			Ethiopia	Malawi	
			AN	CH	
126C, 187T, 189C, 223T, 264T, 270T, 278T, 293G, 311C,	65	9			74
126C, 187T, 189C, 223T, 264T, 270T, 278T, 311C,	38	7		1	46
129A, 209C, 223T, 292T, 295T, 311C,	41	3			44
223T, 278T, 294T, 309G,	37	4		3	44
172C, 183C, 189C, 223T, 320T,	35	7			42
129A, 148T, 168T, 172C, 187T, 188G, 189C, 223T, 230G, 311C, 320T,	31	6	3	1	41
223T, 327T,	28	5		5	38
223T, 265T,	20	12		4	36
124C, 223T, 278T, 362C,	28	5	1	1	35
223T, 320T,	30	2			32
189C, 192T, 223T, 278T, 294T, 309G,	21	10			31
124C, 223T, 278T, 311C, 362C,	24	6			30
148T, 172C, 187T, 188G, 189C, 223T, 230G, 311C, 320T,	2	15	2	11	30
189C, 223T, 278T, 294T, 309G,	28	1			29
172C, 189C, 223T, 320T,	24	2			26
129A, 148T, 168T, 172C, 187T, 188G, 189C, 223T, 230G, 278T, 293G, 311C, 320T,	8	14		3	25
223T, 278T, 286T, 294T, 309G,	15	7		3	25
92C, 223T, 278T, 286T, 294T, 309G,	18	3			21
114A, 129A, 213A, 223T, 278T, 354T,	16	4			20
209C, 223T, 292T, 311C,	20				20
129A, 145A, 187T, 189C, 213A, 223T, 265C, 278T, 286G, 294T, 311C, 360T,	19				19
223T, 278T, 294T, 309G, 368C,	18			1	19
209C, 223T, 311C,	5	13			18

129A, 187T, 189C, 223T, 274A, 278T, 293G, 294T, 311C, 360T,		17			17
223T, 311C, 320T,	16	1			17
129A, 187T, 189C, 223T, 265C, 278T, 286G, 294T, 311C, 360T,	12	4			16
38G, 86C, 129A, 187T, 189C, 223T, 278T, 284G, 293G, 294T, 311C, 360T,	16				16
92C, 223T, 278T, 294T, 309G,	14	2			16
93C, 148T, 172C, 187T, 188G, 189C, 223T, 230G, 311C, 320T,	6	9		1	16
124C, 223T,	15				15
185T, 223T, 311C, 327T,	2	13			15
124C, 183C, 189C, 223T, 278T, 304C, 311C,		14			14
124C, 223T, 256T, 368C,	11	3			14
51G, 223T, 264T,	5	5		3	13
93C, 223T, 265T,	11	2			13
114A, 129A, 213A, 223T, 278T, 355T, 362C,	10	2			12
129A, 223T, 278T, 294T, 309G,	12				12
209C, 218T, 223T, 292T, 311C,	9	3			12
223T, 264T, 278T,	9	3			12
129A, 148T, 172C, 187T, 188G, 189C, 223T, 230G, 311C, 320T,	1	10			11
223T, 278T, 294T,	7	4			11
129A, 187T, 189C, 214T, 223T, 265C, 278T, 286A, 291T, 294T, 311C, 360T,	1	9			10
189C, 192T, 223T, 278T, 294T,	10				10
189C, 192T, 223T, 278T, 294T, 362C,	10				10
93C, 124C, 223T, 278T, 362C,	1	6	1	2	10
Total	749	242	7	39	1037

3.3.6. Conclusion

This study demonstrates that large, anthropologically well defined sample sets from groups in sub-Saharan Africa, along with the right choice of markers, can shed light on past demographic events and the distribution of human genetic diversity in this region. Based on a) typing four informative UEPs within the NRY E1b1a clade in over 2,300 samples from multiple regions, b) analysing frequencies of the haplogroups in the different regions of the continent and STR diversity within the haplogroups and c) estimating TMRCAs of clade-defining UEPs, a pattern was revealed consistent with: a) multiple waves of migration from West Africa southwards, b) a later migration with rapid gene-flow southward along the eastern side of Africa, independent of earlier migration on the western side and c) the absence of substantial east to west NRY gene flow in sub-Saharan Africa over the past millennium. It is noteworthy that estimates of TMRCA for each defining UEP were consistent with the genealogy. Also consistent with the above interpretation is the observation that Nigeria and Cameroon displayed the highest E1b1a STR diversity of any region, as would be expected if this region is, as suggested by linguistic and archaeological evidence, the birthplace of the EBSP.

Analysis of mtDNA HVR-1 haplotypes on the other hand provides no evidence of multiple waves of maternally mediated gene flow (nor does it support an argument that such events did not occur). However, this genetic system does provide evidence of significant inter-regional variation and probable recruitment of local women into the migrating wave of Bantu speakers (based upon observations of mtDNA haplotypes in significant numbers present in Congo (>20%) and Mozambique (10%) that were not seen in a large West African dataset.

4. The *CASP12* stop-codon and its route out of Africa

4.1. Introduction

4.1.1. Evolution of *CASP12* gene

Human caspases are a family of cysteine proteases involved in inflammatory immune response and apoptosis (Fischer et al. 2002; Saleh et al. 2004; Yeretssian et al. 2009). Phylogenetically, the Caspase-12 gene *(CASP12)* has been grouped with the genes encoding caspases-1, 4 and 5 (*CASP1*, *CASP4* and *CASP5*), which are all involved in inflammatory cytokine production. Although its orthologue in mice has been found to be involved in endoplasmic reticulum stress-mediated apoptosis (Nakagawa et al. 2000), the mutation at the third position of the SHG box in human Caspase-12 (G205S) excludes the possibility of such activity (Fischer et al. 2002 and Lamkanfi et al. 2004).

At the genic level, human *CASP12* is found on chromosome 11q22.3 and the highest expression level is found in the lung (Fischer et al. 2002). The longest transcript was found to be 1401bp in length, comprised of eight exons. However, this transcript was only found to be expressed at low levels, while the majority of transcripts were found to be shorter, mainly lacking exon 3. Moreover, all the transcripts missing exon 3 were found to have a premature stop-codon in exon 4 (rs497116), resulting in the formation of a truncated CASP12 protein. Therefore, it was thought that due to this nonsense mutation, *CASP12* was not functional in humans (Fischer et al. 2002). However, further characterisation of this gene in more samples from diverse populations showed that a full length allele (active) also exists and is present at a frequency of 11.4% in individuals with African ancestry from South Africa (n =153) and 10.8% in African Americans (n = 623), whereas European-descent (n = 187) and Asian (n = 160) samples only exhibited the stop-codon allele (Saleh et al. 2004). The frequency distribution of the truncated allele was examined in more population groups within and outside Africa and a similar pattern was observed where the active variant was at high frequency in sub-Saharan Africa and only sporadically found outside

Africa (Kachapati et al. 2006; Xue et al. 2006). In addition, based on diversity indices, it was shown by Xue et al. (2006) that the active variant is the ancestral state of *CASP12* and the stop-codon had occurred on its background rather than the earlier notion that a reverse mutation had abolished the stop-codon allele in some individuals resulting in the full length allele (Kamfanki et al. 2004). Using the phylogeny based time to the most recent common ancestor (TMRCA) method (Bandelt 1999), the SNP creating the stop-codon was estimated to have occurred over 100,000 years ago, prior to the expansion of anatomically modern human out of Africa (Xue et al. 2006). It was also suggested that the worldwide frequency distribution of the stop-codon allele (at fixation outside Africa except for a few Middle Eastern populations and Chinese Han) is mainly due to recent strong positive selection occurring in the Palaeolithic period around 60 KYA.

4.1.2. Phenotypic effects of *CASP12* variation

At the phenotypic level, it has been suggested that individuals with the non-truncated form produce lower levels of cytokines following stimulation by bacterial lipopolysaccharides and thus have a lower innate immune response. This attenuation of the immune response by the active Caspase 12 could result in the development of sepsis (Saleh et al. 2004). This finding was further corroborated by animal studies where CASP12-deficient mice showed greater resistance to bacterial infection (Saleh et al. 2006). Therefore, it was proposed by Xue et al. (2006) that resistance to severe sepsis was the selective advantage associated with the truncated allele. Due to the presence of a single high frequency inactive haplotype (defined by SNPs) in population groups representing sub-Saharan Africa, Europe and Asia, it was suggested that further research might identify the most likely route and date by which the stop-codon polymorphism started to expand from the African continent (Xue et al. 2006).

4.1.3. Aims

The aim of this project is to establish the most likely geographic route taken by the truncated gene in its exit from Africa; in particular, whether it was more likely from

West Africa into Europe or from East Africa into the Middle East, or both are equally likely. The frequencies of the full length and truncated variants of *CASP12* will be determined in multiple groups in both East (n=380) and West (n=373) Africa. If there is no significant difference in the distribution of the stop-codon frequency between East and West, one or more polymorphisms, in linkage disequilibrium (LD) with the stop-codon but variable in frequency between East and West Africa, in the vicinity of the stop-codon will be sought through re-sequencing This will be of great advantage since polymorphisms in complete or high LD with the stop-codon may be differentially present in east, west and outside of Africa and therefore may be used to suggest the most likely route taken by the truncated allele out of Africa.

4.2. Methods

4.2.1. Samples

Buccal swab samples from males over eighteen years old, unrelated at the paternal grandfather level but otherwise randomly selected from twelve sub-Saharan African, two North African and two Middle Eastern groups, were used to type the *CASP12* stop-codon SNP (rs497116). In sub-Saharan Africa, seven groups represented West Africa and areas covered by the expansion of Bantu-speaking peoples (West Africa/EBSP ascertainment panel; n =373) and five groups represented Ethiopia (Ethiopian ascertainment panel; n = 380). The West Africa/EBSP ascertainment panel comprised the following population groups: Mozambique (Sena; n = 51), Ghana (Asante and Bulsa; n = 57), Malawi (Chewa; n = 50), Congo Brazzaville (Bakongo; n = 55), Cameroon Grassfields (Mambela; n = 65), Lake Chad (Shewa Arabs; n = 65) and Sudan (Kordafanians; n = 30). The Ethiopian ascertainment panel comprised five ethnic groups namely Afar, Amhara, Oromo, Anuak and Maale with equal sample sizes (n = 76). The two North African groups were from Morocco (n = 48) and Algeria (n = 48), and the two Middle Eastern groups were from Yemen (n = 48) and Iran (n =48).

Introns surrounding exon 4 of *CASP12* were resequenced in buccal swab DNA samples collected from Oromo and Amhara (n=95), representing East Africa, Mozambique and Ghana (n = 95), representing West Africa and EBSP. In order to obtain a better distribution of *CASP12* haplotypes outside Africa, another eight groups representing North Africa (Morocco and Algeria), the Middle East (Iran and Yemen), Europe (Wales, Friesland and Ukraine) and Asia (China) (equal sample size of 48; total of 384) were also re-sequenced for intron 4 only. This choice was made during this study based on the observation that this intron carried a short tandem repeat (STR) polymorphism with differential frequency distribution between populations representing East and West Africa.

4.2.2. Genotyping

Tetra-primer ARMS PCR (Ye et al. 2001) with minor modifications was used to genotype the stop-codon SNP. The outer and two inner fragments were amplified in a 10-µl reaction volume containing 1 µl (~ 1 ng) of template DNA, 1.5 pg of each outer primer, 7.5 pg of each allele-specific primer (inner primers), 1.6 µl (50 µM) dNTPs, 9.3 nM TaqStartmonoclonal antibody (BD Biosciences Clontech, Oxford, UK), 1 µl of 10x Taq buffer and 0.13 units of Taq DNA polymerase (HT Biotech, Cambridge, UK) (see Table 4.1 for primer details). All samples (96-well plates) were then placed on a thermocycler under the following conditions: denaturation at 95^{o}C for 5 min, followed by 35 cycles of 95^{o}C for 45 s, 61^{o}C for 45 s and 72^{o}C for 45 s. The final step of the PCR program was a 7-min extension at 72^{o}C, before a 30-min hold at 4^{o}C. PCR bands were run on a 2% (w/v) agarose gel and visualised under UV light.

Table 4.1. Details of the primers used in the tetra-primer ARMS PCR reaction

Primer name	Primer sequence (5'-3')	Melting temperature (^{o}C)	GC %	Fragment size (bp)	Primer concentration (µM)
CASP12-EX4-OUT-F	ACGAGGGTGTATTTTCATGCAG	56.4	45.5	401	0.15
CASP12-EX4-OUT-R	CTTGCTCTTTCAGCTGCCAA	57.2	50.0		0.15
CASP12-T-ALLELE	AACTTGACCTTTTGGGGAGGT	57.5	47.6	179	0.75
CASP12-C-ALLELE	CAAGGTTTTCAAGTAGATCGCG	55.3	45.5	264	0.75

4.2.3. Re-sequencing

DNA resequencing was undertaken to identify polymorphisms in the vicinity of the stop-codon in the following manner. Two sets of primers were designed to amplify two overlapping PCR fragments with the first covering 690 bp of intron 3 flanking exon 4 (690 bp) and the second fragment (739 bp) including all exon 4 (173 bp) and 437 bp of flanking region of intron 4 (see Table 4.2 for primer details). Each fragment was amplified separately in 10-µl reaction volumes containing 1 µl (~ 1 ng) of template DNA, 1.5 pg of each primer (forward and reverse), 1.6 µl (50 µM) dNTPs, 9.3 nM TaqStartmonoclonal antibody (BD Biosciences Clontech, Oxford, UK), 1 µl of 10x Taq buffer and 0.13 units of Taq DNA polymerase (HT Biotech, Cambridge, UK). The reaction mixtures (96-well plates) were incubated at 95°C for 4 min, followed by 39 cycles of 95°C for 1 min, 61°C for 1 min and 72°C for 1 min. The final step of the PCR program was a 7 min extension at 72°C before a 30 min hold at 4°C.

Table 4.2. Primer details for the PCR fragments amplifying exon 4 and flanking introns in *CASP12*.

PCR fragment	Primer Name	Sequence (5' - 3')	Length (bp)	Tm (°C)	GC %
Intron 3 (690 bp)	CASP12-INT3-F	CCAGGTCAGTAAGAAGCAGAAGGA	24	58.9	50.0
	CASP12-INT3-R	CACAGAATGACTTTCCCCAGGA	22	57.8	50.0
Exon 4-Intron 4 (739 bp)	CASP12-INT4-F	ACGAGGGTGTATTTTCATGCAGAA	24	57.4	41.7
	CASP12-INT4-R	GTGTGTGTGATTTGTTCCCCCTA	23	58.1	47.8

PCR products were purified by mixing 30 µl of a 1 to 2 mix of water and HM-MC (40% PEG- 8000, 1 M NaCl, 2 mM Tris-HCl (pH 7.5), 0.2 mM EDTA, 3.5 mM $MgCl_2$) to each sample. The mixtures were then centrifuged at 2240 g for 45 min, followed by centrifuging the inverted PCR plate for 1 min at 13x g in order to discard the supernatant. 150 µl of 70 % ethanol was added to each sample and the plate was centrifuged at 2240 x g for 25 min. The resultant supernatant was discarded by the same method as above. Samples were then dried on a thermocycler at 65°C for 5 min followed by eluting each sample with 30 µl of distilled water. Finally, the eluted samples were mixed gently to resuspend the DNA pellet.

The same forward and reverse PCR primers were used for resequencing in both directions. PCR fragments were sequenced in 10-µl reaction volumes containing 2 µl of the purified PCR product, 0.16 µM of sequencing primer, 0.35 µl of BigDye termination mix v3.1 (Applied Biosystems) and 2.15 µl of the accompanied 5x sequencing buffer. The samples were placed on a thermocycler under the following conditions: 25 cycles of 96°C for 10 seconds, 55°C for 5 seconds and 60°C for 4 minutes.

To purify the sequencing reaction products, 2.5µl of 125 mM EDTA and 30 µl of 100% ethanol were added to each sample, thoroughly mixed and left at room temperature for 10 min. The mixture was centrifuged for 60 min at 2240 x g, followed by centrifuging the inverted PCR plate at 13x g for 1 min to discard the supernatant. 30 µl of 70 % ethanol was then added to each sample and centrifuged for 10 minutes at 2240 x g. The resultant supernatant was again discarded by inverting the PCR plate and centrifuging for 1 min at 13x g. Samples were then dried for 5 minutes at 65 ºC before adding 10 µl of Hi-Di formamide prior to electrophoresis. Finally, samples were run on an ABI 3730 genetic analyser and analysed using Sequencher 4.7 software (Gene Codes Corporation, USA). The chromatogram of each sample was visually inspected across its entire length to check for high levels of background noise that would make it difficult to call bases. Subsequently, if no significant trace of noise was found, the ends of the sequence were trimmed down to the reference sequence (usually 50bp after the start of the PCR fragment in the direction sequenced). A contig was then formed for each set of 96 samples (i.e., one sequencing run) and each mutation site examined visually to determine whether it was a substitution, insertion, deletion or a tandem repeat. All samples with any ambiguous sites were resequenced.

4.2.4. Statistical analysis

Haplotypes were inferred by parsimony and checked using the Excoffier-Laval-Balding (ELB) and Expectation Maximisation (EM) algorithms (Excoffier et al. 2003 and Excoffier & Slatkin 1995 respectively). Pairwise linkage disequilibrium (LD) was assessed between loci by estimating D´. Gene diversity, *h,* and its standard error were estimated from unbiased formulae of Nei (1987). Patterns of genetic differentiation were quantified using F_{ST} (Reynolds, Weir & Cockerham 1983) estimated from AMOVA-based φ-st values (Excoffier et al. 1992 and Michalakis & Excoffier 1996). Significance of genetic distances was assessed by permutation test where a null distribution is formed by calculating all possible values under rearrangements of haplotypes in the observed samples. The permutation test was repeated over 1,000 times to give the null distribution. All the above was performed using Arlequin software version 3.0 (Excoffier et al. 2005).

Using MS Excel, deviation from Hardy-Weinberg expectations was assessed at each locus and its significance was examined based on the Chi-squared distribution. A two-tailed Z test was also undertaken in MS Excel to test significance of gene diversity differences. Principal Coordinates Analysis (PCO) was performed using the 'R' environment of statistical computing (www.R-project.org) by implementing the 'cmdscale' function on pairwise F_{ST} matrices and visualised using the 'plot' function. The plot was used to visualise relationships among groups. Average Square Distances (ASD) were also calculated in MS Excel and corresponding 95% confidence intervals were calculated as in Thomas et al. (1998) using 'R'. The Time to Most Recent Common Ancestor (TMRCA) were estimated using a STR mutation rate of 1.94×10^{-4} for a dinucleotide STR (Huang et al. 2002) and generation time of 25 years. Relationships of *CASP12* haplotypes were displayed via a reduced-median network constructed within Network 4.56 (Fluxus Engineering).

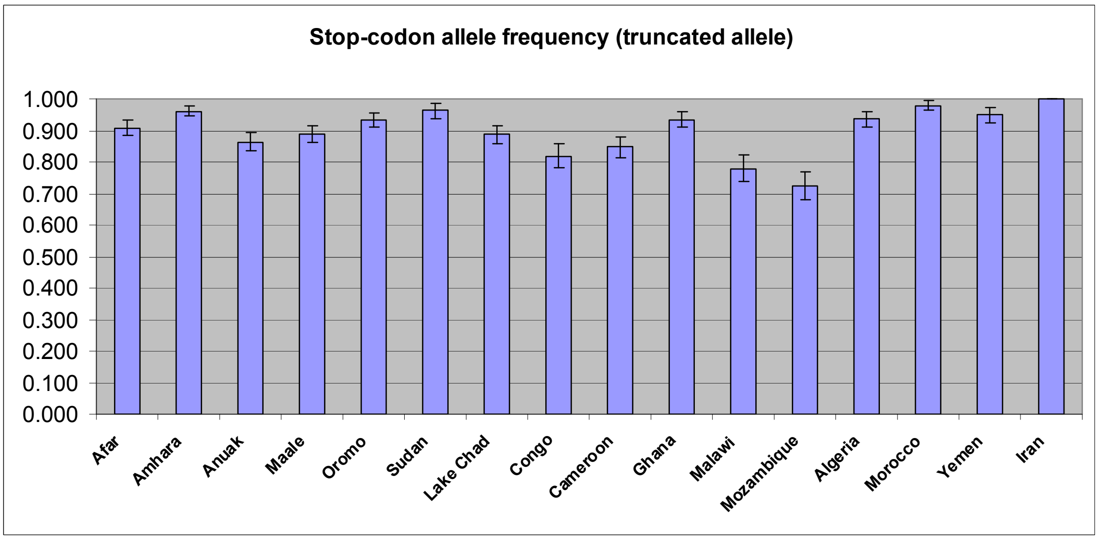
4.3. Results and discussion

4.3.1. Is there difference in the prevalence of the nonsense mutation between populations in West Africa and East Africa?

The prevalence of the nonsense mutation was assessed in all samples as described above. The nonsense mutation is present at high frequency (see Figure 4.1) across all sub-Saharan African populations. Deviation from Hardy-Weinberg equilibrium

(HWE) was tested, with only the Afar indicating deviation from it (p >0.125, Afar p=0.0012).

Figure 4.1. The distribution of the truncated *CASP12* gene variant in five Ethiopian ethnic groups and eleven countries.



Note: The error bars here represent the standard deviation of each allele frequency

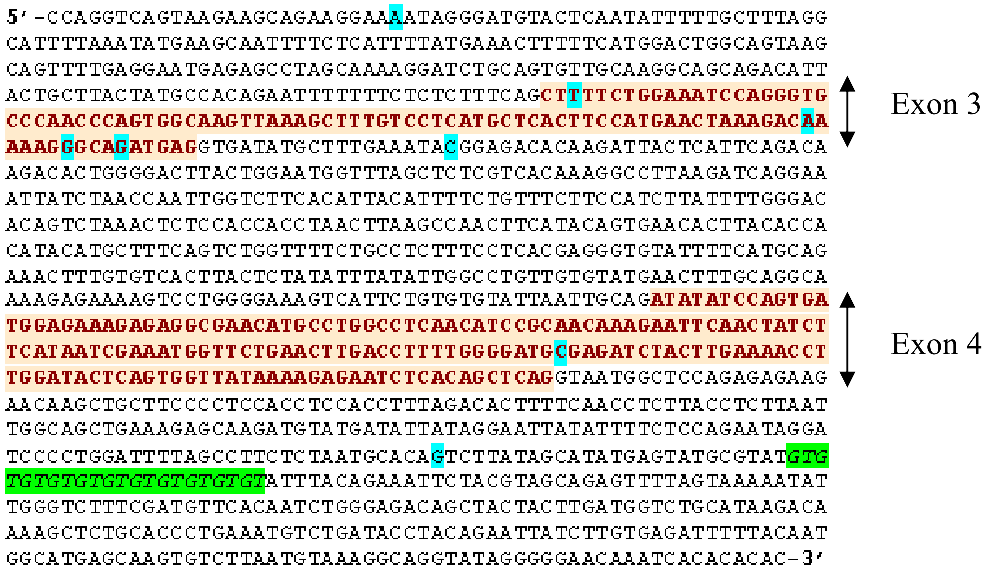
The deviation from HWE observed in Afar remained significant after applying a Bonferroni correction. It is possible that the deviation arises from abrogation of one or more assumptions made in formulating HWE expectations, in this case quite possibly that of random mating given the reported high level of consanguineous unions among the Afar (Getachew 2001). Based on the stop-codon allele frequencies, there was no significant difference between the West African (i.e., Ghana and Cameroon) and Ethiopian groups (Fisher's exact test p = 0.355).

4.3.2. Is there variation in the genomic regions flanking the nonsense mutation?

Given that there was no statistically significant difference in the prevalence of the nonsense mutation between West and East African groupings, the genomic regions flanking the nonsense mutation were examined to identify variations which might be used to construct haplotypes (that included the nonsense mutation) that were at different frequencies in the two regions. Figure 4.2 shows the region resequenced (1300 base pairs in total) in 43 Ghanaian samples representing West African populations and 76 Oromo samples representing Ethiopia. Figure 4.2 and Table 4.3

detail the variant sites and their frequencies. Seven variants were single base substitutions and one was a variable short tandem repeat (STR). Of the single base substitutions, three were novel and, of these previously unreported alleles, one was only observed in a single individual.

Figure 4.2. The graphical representation of the resequenced genomic region flanking the stop-codon in exon 4 of *CASP12* gene (human assembly GRCh37). Exonic sequence is shown in red and identified polymorphisms are highlighted in blue (SNPs) and green (STR).



```
5'-CCAGGTCAGTAAGAAGCAGAAGGAAAATAGGGATGTACTCAATATTTTTGCTTTAGG
CATTTTAAATATGAAGCAATTTTCTCATTTTATGAAACTTTTTCATGGACTGGCAGTAAG
CAGTTTTGAGGAATGAGAGCCTAGCAAAAGGATCTGCAGTGTTGCAAGGCAGCAGACATT
ACTGCTTACTATGCCACAGAATTTTTTTCTCTCTTTCAGCTTTTCTGGAAATCCAGGGTG
CCCAACCCAGTGGCAAGTTAAAGCTTTGTCCTCATGCTCACTTCCATGAACTAAAGACAA
AAAGGGCAGATGAGGTGATATGCTTTGAAATACGGAGACACAAGATTACTCATTCAGACA
AGACACTGGGGACTTACTGGAATGGTTTAGCTCTCGTCACAAAGGCCTTAAGATCAGGAA
ATTATCTAACCAATTGGTCTTCACATTACATTTTCTGTTTCTTCCATCTTATTTTGGGAC
ACAGTCTAAACTCTCCACCACCTAACTTAAGCCAACTTCATACAGTGAACACTTACACCA
CATACATGCTTTCAGTCTGGTTTTCTGCCTCTTTCCTCACGAGGGTGTATTTTCATGCAG
AAACTTTGTGTCACTTACTCTATATTTATATTGGCCTGTTGTGTATGAACTTTGCAGGCA
AAAGAGAAAAGTCCTGGGGAAAGTCATTCTGTGTGTATTAATTGCAGATATATCCAGTGA
TGGAGAAAGAGAGGCGAACATGCCTGGCCTCAACATCCGCAACAAAGAATTCAACTATCT
TCATAATCGAAATGGTTCTGAACTTGACCTTTTGGGGATGCGAGATCTACTTGAAAACCT
TGGATACTCAGTGGTTATAAAAGAGAATCTCACAGCTCAGGTAATGGCTCCAGAGAGAAG
AACAAGCTGCTTCCCCTCCACCTCCACCTTTAGACACTTTTCAACCTCTTACCTCTTAAT
TGGCAGCTGAAAGAGCAAGATGTATGATATTATAGGAATTATATTTTCTCCAGAATAGGA
TCCCCTGGATTTTAGCCTTCTCTAATGCACAGTCTTATAGCATATGAGTATGCGTATGTG
TGTGTGTGTGTGTGTGTGTGTATTTACAGAAATTCTACGTAGCAGAGTTTTAGTAAAAATAT
TGGGTCTTTCGATGTTCACAATCTGGGAGACAGCTACTACTTGATGGTCTGCATAAGACA
AAAGCTCTGCACCCTGAAATGTCTGATACCTACAGAATTATCTTGTGAGATTTTTACAAT
GGCATGAGCAAGTGTCTTAATGTAAAGGCAGGTATAGGGGGAACAAATCACACACAC-3'
```

Table 4.3. Details of *CASP12* polymorphisms identified in this study.

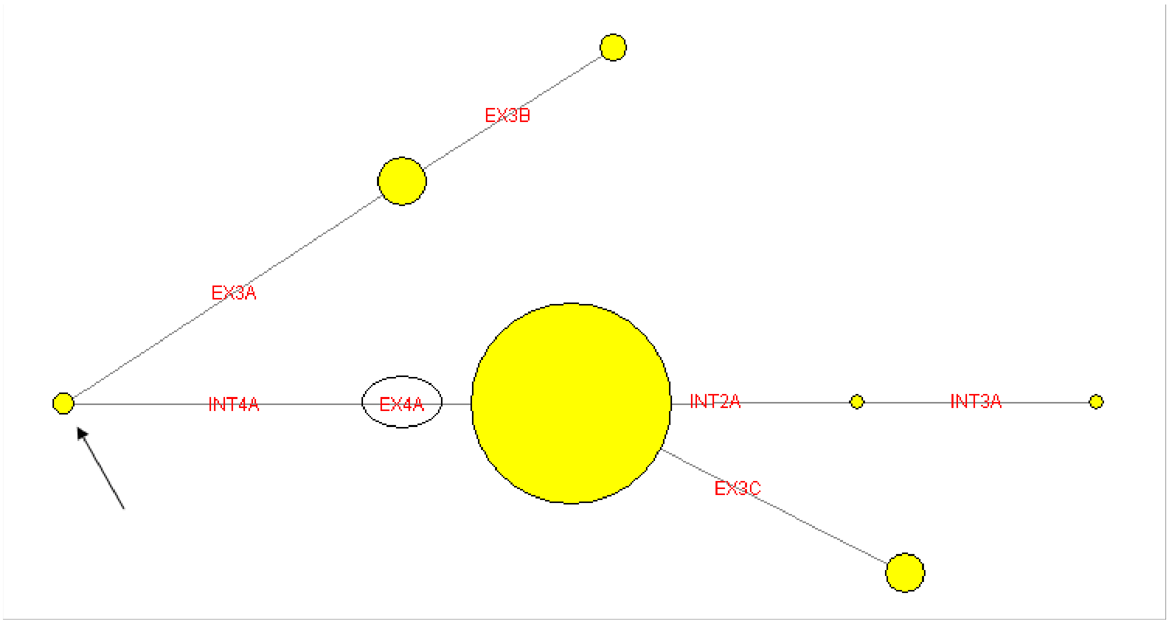
Name of variant	Type of variant	Nucleotide change	dbSNP ID	Mutation type	MAF (Ethiopia)	MAF (Ghana)
INT2A	SNP	A/G	Novel		0.000	0.027
EX3A	SNP	C/T	rs617983	Missense	0.070	0.054
EX3B	SNP	A/G	Novel	Sense	0.023	0.000
EX3C	SNP	G/T	rs73629659	Missense	0.016	0.068
INT3A	SNP	C/G	Novel		0.000	0.014
EX4A	SNP	C/T	rs497116	Nonsense	0.078	0.081
INT4A	SNP	G/T	rs648264		0.078	0.081
INT4B	STR	$(GT)_{9-16}$	Novel		NA**	NA

* SNP ID was obtained from the NCBI SNP database (**http://www.ncbi.nlm.nih.gov/SNP**) **NA: Not Applicable

4.3.3 Is there evidence of recombination in the genomic region that includes the nonsense mutation and its flanking regions?

To examine whether recombination could be detected in the region which includes the stop-codon, a phylogeny-based method (Hudson & Kaplan 1985; Maynard Smith & Smith 1998) was implemented. This method works on the basis that, if there is neither recombination nor homoplasy, the number of variable sites is equal to the number of steps in a most-parsimonious tree. By adopting parsimony, a genealogical tree of all single base substitutions identified was constructed (see Figure 4.3).

Figure 4.3. Genealogical tree of single-base substitutions identified in the re-sequenced region. Nodes are proportional to the frequencies of each sequence. The arrows show the ancestral sequence and the circled marker is the nonsense mutation step.



The number of variable sites was equal to the number of tree steps. Moreover, the tree produced is the only possible genealogical relationship among the identified variable sites. Therefore, it is possible to assume that no recombination has taken place in this region. Furthermore, linkage disequilibrium patterns observed in this region also provided no evidence of recombination (D′ of 1 among all loci). Consequently, distributions of haplotypes carrying the stop-codon were examined between East and West Africa.

4.3.4 Given that there is no evidence of recombination what is the estimated prevalence of haplotypes (including nonsense mutation STR haplotypes) in the African and non-African groups?

Estimated haplotype frequencies based upon single nucleotide substitutions are included in Table 4.4, while Table 4.5 contains frequencies of all STR sizes of chromosomes homozygous for the nonsense allele (T dataset).

Table 4.4. Frequencies of inferred SNP haplotypes, based on parsimony and the ELB algorithm, in Ghana and Ethiopia. Haplotypes 1-3 carry the ancestral allele (underlined) at the stop-codon SNP.

Haplotype number	Haplotype sequence	Ghana	Ethiopia	Total
1	ACGGC$\underline{C}$G	0.027	0.008	0.015
2	ATGGC$\underline{C}$G	0.054	0.047	0.050
3	ATAGC$\underline{C}$G	0.000	0.023	0.015
4	ACGGCTT	0.824	0.906	0.876
5	ACGTCTT	0.068	0.016	0.035
6	GCGGGTT	0.014	0.000	0.005
7	GCGGCTT	0.014	0.000	0.005

Gene diversity was calculated based on STR alleles observed in the T dataset and ranged in individual populations from 0.316-0.641, where Wales and Ghana had the lowest and highest values. Samples were then pooled into African and non-African datasets and gene diversities were calculated. Diversity of the African dataset (h = 0.557) was higher than the non-African dataset (h = 0.374) and the difference was highly significant (P < 0.0001). The significant reduction of genetic diversity from Africa to outside is consistent with the diversity pattern of other nuclear loci investigated (Tishkoff et al. 2009).

Table 4.5. The distribution of *CASP12* truncated haplotypes in multiple geographic locations. Haplotypes are defined by the stop-codon allele (T) and the novel STR alleles (range of 10-16) in intron 4.

Geographic location	*CASP12* truncated haplotypes						
	T 10	T 11	T 12	T 13	T 14	T 15	T 16
Ghana (GHA)		0.186	0.443	0.371			
Ethiopia (ETH)	0.008	0.042	0.691	0.225	0.017	0.017	
Iran (IRA)		0.011	0.793	0.011	0.022	0.163	
Yemen (YEM)		0.050	0.800	0.063	0.063	0.024	
Algeria (ALG)		0.011	0.649	0.181	0.096	0.064	
Morocco (MOR)		0.011	0.644	0.133	0.156	0.056	
China (CHI)		0.094	0.690	0.024	0.192		
Wales (WAL)			0.817		0.049	0.134	
Friesland (FRI)			0.766		0.106	0.128	
Ukraine (UKR)			0.819	0.021	0.074	0.074	0.011

4.3.5. Using STR variation to calculate genetic distances

There was no significant difference between the distributions of nucleotide substitution defined haplotypes between West and East Africa ($F_{ST} < 0.001$, $P = 0.445$). The same two geographic regions, based on the distribution of STR alleles (using just the nonsense mutation bearing chromosomes for which STR status could be unambiguously assigned), were compared using pairwise genetic distances. Pairwise R_{ST} was estimated to be < 0.001 ($P = 0.618$), while F_{ST} was 0.083 and highly significant ($P = 0.002$). Since genetic differentiation between East and West Africa was obtained only using F_{ST} based on the STR alleles shown in Table 4.5, all other comparisons among the ten populations typed for the stop-codon and the STR marker were undertaken in the same manner (see Table 4.6). Based on F_{ST} values, one cluster was apparent, which comprised of the three European populations along with Iran and Yemen ($F_{ST} \leq 0.024$, $P > 0.05$).

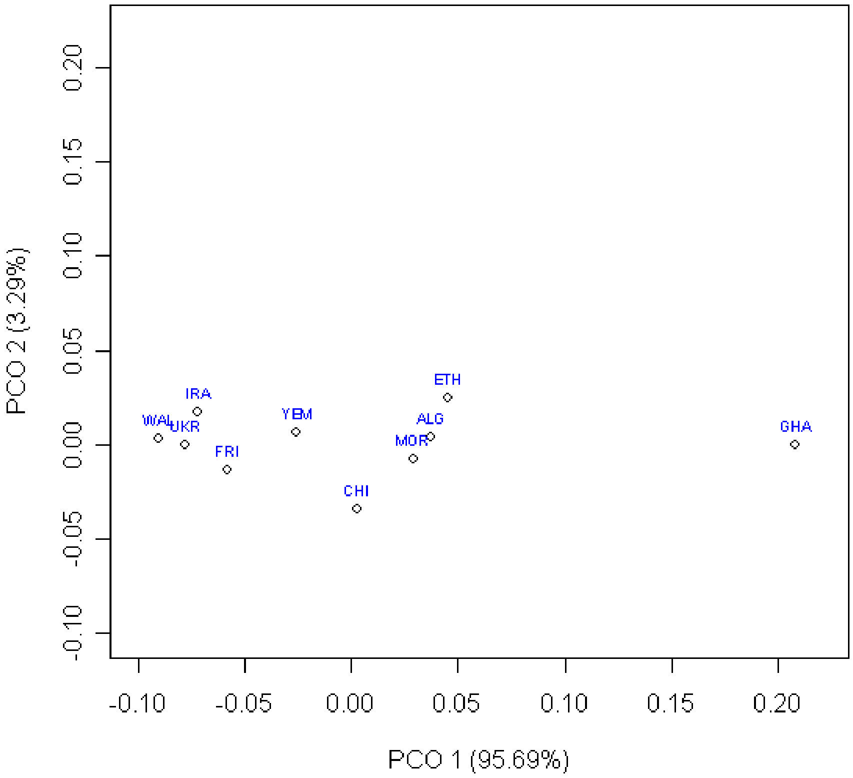
Table 4.6. Pairwise genetic distances (Fst) based on STR alleles restricted to the T dataset in ten populations.

	Ghana	Ethiopia	Iran	Yemen	Algeria	Morocco	China	Wales	Friesland
Ethiopia	0.083								
Iran	0.274	0.081							
Yemen	0.214	0.036	**0.024**						
Algeria	0.089	**0.004**	0.061	0.033					
Morocco	0.113	0.022	0.062	0.032	**<0.001**				
China	0.175	0.065	0.074	0.026	0.034	0.013			
Wales	0.296	0.087	**<0.001**	**0.016**	0.065	0.061	0.065		
Friesland	0.258	0.077	**0.002**	**0.015**	0.043	0.032	0.033	**<0.001**	
Ukraine	0.283	0.070	**0.007**	**<0.001**	0.052	0.046	0.044	**<0.001**	**<0.001**

Significant pairwise genetic distances (P < 0.05) are shown in bold red.

The PCO plot in Figure 4.4 enables the relative positions of the groups to be visualised. Ghana, representing West Africa, is clearly at a greater distance from the non-African groups than is Ethiopia.

Figure 4.4. PCO plot of pairwise Fst genetic distances based on *CASP12* truncated haplotypes



Percentages in parentheses represent the amount of variation explained by each eigen vector.

The greater similarity of non-African groups to Ethiopia could be the consequence of back migration into Africa from the Arabian Peninsula or elsewhere, either associated with the migration of Semitic speakers some 5,000 years ago (Cavalli-Sforza et al. 1994), or some other event. However, the smaller distance between Ethiopians and North Africans than between Ethiopians and Yemeni argues against this explanation, leaving the suggestion that the nonsense mutation is more likely to have taken a route out of Africa via the East (represented in this study by Ethiopians) than the West (represented in this study by Ghanaians).

4.3.6 Using variation in the STR, is the estimated age of the nonsense mutation consistent with the mutation occurring prior to the migration of Anatomically Modern Human out of Africa?

If the nonsense mutation predated the mass migration of Anatomically Modern Human out of Africa it would be expected to have an age of at least 60000 years (Quintana-Murci 1999; Campbell & Tishkoff 2010). To estimate its age, Average Squared Distance (ASD) was calculated for the STR alleles of the nonsense mutation containing chromosomes to which the STR size could be unambiguously assigned (i.e., homozygous pairs of chromosomes) to estimate a time to The Most Recent Common Ancestor (TMRCA). To avoid complications associated with any bottleneck during the migration out of Africa (as evident by the significantly reduced STR gene diversity outside Africa), only African samples were used in the calculation. The observed ASD value was 0.864 and the 95% confidence interval was 0.717-1.029. Assuming a mutation rate of 1.94 x 10^{-4} (Huang et al. 2002) and a generation time of 25 years, the nonsense mutation can be dated back to 92400-132600 years ago, a date which is consistent with it having occurred shortly before the exodus out of Africa.

4.3.7. Conclusion

Overall, it has been shown like many previous studies (Kachapati et al. 2006 and Xue et al. 2006) that the truncated *CASP12* variant is at high frequency in many parts of Africa and is almost fixed outside Africa. The analysis in this study suggests that the truncated variant most likely left Africa via an eastern route into Eurasia. This is in good agreement with other studies suggesting a major eastern route of dispersal of

Anatomically Modern Human out of Africa, based on mtDNA (Quintana-Murci et al.. 1999) and nuclear markers (Tishkoff et al. 2009). The estimated age of the truncated variant is consistent both with the wide distribution of the truncated variant in sub-Saharan Africa and its origin prior to the migration out of Africa.

5. Analysing variation in *CYP1A1* and *CYP1A2*: relevance for healthcare, human evolution and anthropology in Africa

5.1. Introduction

5.1.1. Drug-metabolising Enzymes

In nature, organisms, including humans, have evolved complex systems that inactivate and detoxify foreign chemicals (i.e., xenobiotics) including dietary constituents (Rang and Dale 2006, Buxton 2006). In general, four processes are responsible for the rate at which a xenobiotic compound passes through the human body (Weinshilboum 2003). The first step is absorption, when the molecule is absorbed into the bloodstream. The second step is the distribution of the molecule to the site of action where it interacts with target molecules. The third step is metabolism (i.e. biotransformation), where the structure and properties of the ingested molecule are altered. (Although sometimes this alteration could result in bioactivation, it is usually in the form of either inactivation or water solubilisation (Buxton 2006)). The final step is excretion, where the metabolised molecule is removed from the body through renal elimination.

Although genetic variation may influence all four processes, since they involve many different proteins encoded by genes, metabolism has received the most attention. Hitherto, more than 30 gene families have been indentified which are involved in metabolising xenobiotics including drugs (Evans and McLeod 2003). Historically, drug metabolising enzymes (DME) have been divided into two groups (phase I and II), based on the type of their reactions (Weinshilboum 2003). Phase I reactions are catabolic (e.g., oxidation, reduction and hydrolysis), whereas phase II reactions are anabolic (e.g., acetylation and methylation) and involve conjugation (Rang and Dale 2006). Major enzymes in phase I reactions are cytochrome P450 monooxygenases (CYP), flavin-contatining monooxygenases (FMO) and epoxide hydrolases (EH). In phase II metabolism, several enzyme families are also known and the more important among them are UDP-glucuronosyltransferases (UGT), glutathione-*S*-transferases

(GST), sulfotransferases (SULT), *N*-acetyltransferases (NAT), and methyltransferases (MT) (Buxton 2006).

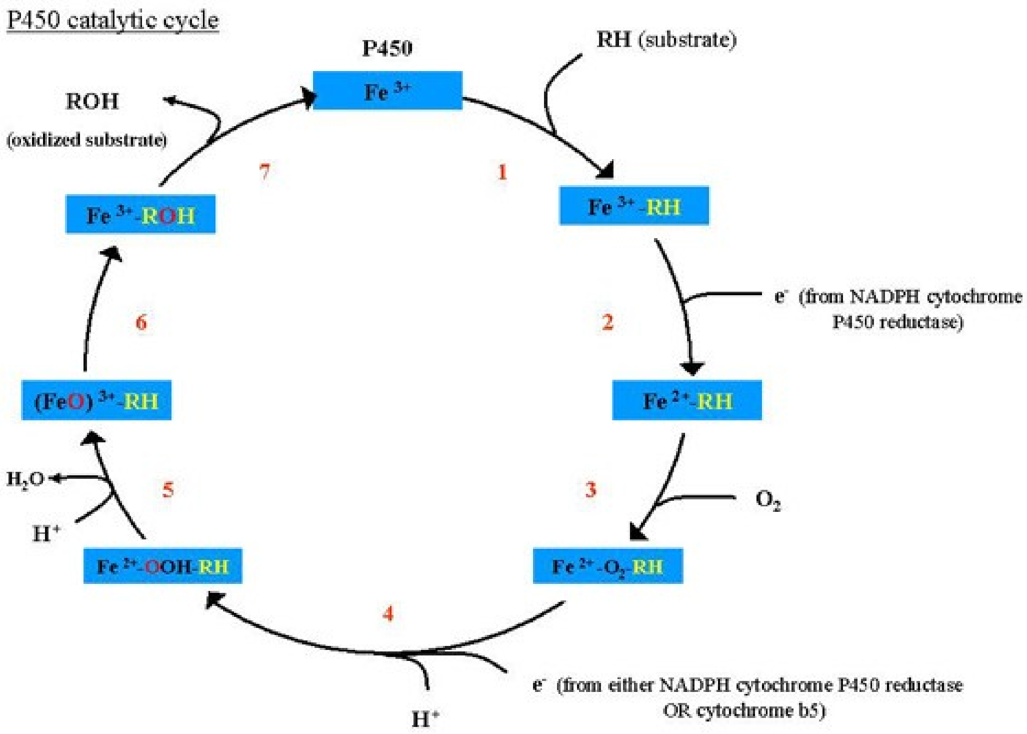
5.1.2. CYP Superfamily

Cytochrome P450 enzymes (CYPs for short) are a superfamily of diverse but related proteins, all of which contain a haem group that is non-covalently attached to the polypeptide chain (Rang and Dale 2006). They are believed to be the most important of the enzymes that catalyze phase I drug metabolism (Weinshilboum 2003).These enzymes were first found in rat liver microsomes by Klingenberg (1958). When combined with carbon monoxide, they showed a characteristic intense absorption band at 450nm when treated with dithionite and hence the name P450. However, it was a few years later that the nature of these enzymes (referred to as 'pigments') as novel cytochrome hemeproteins was identified (Omura and Sato 1962; 1964).

Their function was also characterised and they were shown to be involved in the hydroxylation of several steroids and drugs (Porter and Coon 1991). It should be noted that the function of CYPs is always coupled with another enzyme known as NADPH-dependent P450 reductase (Rang and Dale 2006). Therefore, the lack of this enzyme, which supplies the electron and H^+ required by CYPs to oxidise substrates, is detrimental to CYP activity (see steps 2 and 4 in Figure 5.1). It is now known that beside being powerful *in vivo* oxidising agents, able to catalyse the oxidative biotransformation of a wide range of chemically and biologically unrelated exogenous and endogenous substrates (Porter and Coon 1991), CYPs can also catalyse many other reactions including *N*-dealkylation, *O*-dealkylation, deamination, and dehalogenation (Gonzalez and Tukey 2006)

CYPs have been identified across all kingdoms of life including animals, plants, fungi, protists and bacteria (Nelson 2009). Based on the identification of more than 20 CYPs in bacterial strains, it is thought that cytochromes P450 arose before the split of bacteria and eukaryotes (Nelson 1999). This has further led to the suggestion that CYPs may have existed throughout nature for more than three billion years (Nebert and Russell 2002) and have expanded via divergent evolution (Nebert and Gonzalez 1987).

Figure 5.1. Schematic representation of the CYP450 catalytic cycle



(Figure taken from http://www.reactome.org/figures/p450_cat_cycle.JPG)

The number of putatively functional CYPs varies across species, especially among the eukaryotes with the range of 2-323 where the lowest and highest were found in yeast and plant species respectively (Nelson 2004). The first systematic standardised nomenclature for a protein superfamily was made for CYPs, which was based on phylogenetic relationships (Nebert et al. 1987; Nebert and Gonzalez 1987). The criteria used to classify CYP proteins into families and subfamilies is amino-acid sequence identity. Currently, enzymes that share ≥40% amino-acid identity are assigned to a particular family (e.g., CYP1), whereas those sharing ≥55% identity make up a particular subfamily (e.g., CYP1A) (Nebert and Russell 2002).

In terms of function, different CYPs have distinct, but often overlapping, substrate specificities, with some enzymes acting on the same substrates as each other but at different rates (Rang and Dale 2006). The structural and functional data on CYPs suggest that enzymatic conformational flexibility may be central to the ability of CYPs to bind a diverse array of substrates (Coon 2005) and therefore at times have overlapping functions.

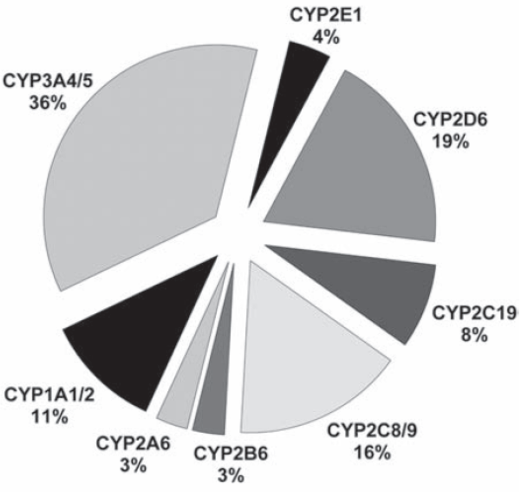
5.1.3. Human CYP and Pharmacogenetics

In humans, 57 putatively functional genes (18 families and 42 sub-families) along with numerous pseudogenes have been identified (Nelson 2004). This number of functionally related but distinct genes results in the synthesis of a variety of CYP isozymes that are able to catalyse the metabolism of widely different chemical structures. It has been estimated, therefore, that CYP substrates number over 200,000, which include several types of endobiotics and many xenobiotics (Urban et al. 2001). Endogenous substrate classes identified include fatty acids, eicosanoids, steroids, bile acids, retinoids and vitamin D3 derivatives, whereas exogenous substrates include pharmaceutical drugs, environmental pollutants and plant products (Nerbert and Russell 2002). CYPs were originally thought to be a hepatic system. However, it is now known that CYPs are also expressed in many other tissues, albeit to a lesser extent, including lungs and the intestine (Ingelman-Sundberg 2004a).

Factors such as age, sex, disease state and drug interactions result in variable drug response in individuals and variation in CYP isozymes can also influence drug response significantly (Pirmohamed and Park 2004). Therefore, pharmacogenetics, which is the study of differential drug response based on genetic variation in individuals (Weinshilboum 2003), is of considerable importance. It is thought that CYP isozymes catalyse 75% of phase I metabolism of pharmaceutical drugs with CYP families 1-3 being mostly responsible. Interestingly, genes in CYP families 1-3 are less conserved evolutionary compared with CYP families 5-51 and display important genetic polymorphisms (Ingelman-Sundberg 2004a). These polymorphisms may lead to abolished, reduced, altered and increased activity of the encoded enzyme (Pirmohamed and Park 2004). Furthermore, since the biotransformation of xenobiotics by CYPs can lead to toxicity by metabolic activation of procarcinogens and drugs (Ingelman-Sundberg 2004a), polymorphisms affecting the rate of expression are also of key importance.

Polymorphic CYPs have also been associated with the development of a significant number of adverse drug reactions (ADRs) (Ingelman-Sundberg 2004b). This is due to the observation that about 60% of ADR-related drugs are metabolised by phase I polymorphic enzymes and 86% of those belong to the CYP superfamily (Phillips et al. 2001). It is now generally recognised that ADR is an important issue in both drug treatment and drug development. A study by Lazarou et al. (1998) showed that serious ADRs occur in 7% of hospitalised patients and around 0.3% develop fatal ADRs resulting in around 100,000 deaths a year in the United States of America (US) alone. Another study has shown that the cost of ADRs is an estimated $100 billion per year in the US (Marshall 1997). Interestingly, similar to the position in the US, a pilot study found that around 8% of all hospital admissions in UK were also due to ADRs (Green et al. 2000). Detailed characterisation of genetic variation in CYP genes would be one way forward to tackle this problem since drugs could be administered based on the genotype of an individual or a common variant in a population group. Consequently, pharmacogenetics of CYPs has the potential to identify the right dose for each patient without the need to eliminate drugs based on their pharmacokinetic properties (Ingelman-Sundberg 1999).

Figure 5.2. Proportion of phase I pharmaceutical drug metabolism by CYP isozymes



Adapted from Mechanisms of Drug Interactions I in Drug interactions of infectious diseases by Kashuba and Bertino (2005) Humana Press.

As mentioned earlier, CYP 1-3 families are mostly responsible for drug metabolism (see Figure 5.2). However, just a few sub-families (CYP1A, CYP2C, CYP2D and CYP3A) cover most of this metabolism, with CYP1A metabolising around 11% of drugs, with only two functional genes (Kashuba and Bertino 2005). The biochemistry and genetics of CYP1A proteins and corresponding genes will be dealt with in more detail below.

5.1.4. CYP1A Subfamily

The human CYP1A enzyme subfamily is comprised of two members, CYP1A1 and CYP1A2. These two enzymes have ~70% amino-acid sequence identity (Urban et al. 2001; Moorthy 2008) and the genes encoding these two enzymes form a cluster in a region spanning 39kb on human chromosome 15q22ter (Corchero et al. 2001). It is thought that these two genes originated around 250 million years ago (Nebert and Gonzalez 1987; Heilmann et al. 1988) after the evolutionary split of sea and land animals and that *CYP1A2* came into existence via duplication of *CYP1A1* and further diverged (Nelson et al. 1996).

The expression of *CYP1A* genes is induced by polycyclic aromatic hydrocarbons (PAHs) (e.g., benzo[a]pyrene), β-naphthoflavone (BNF) and other compounds (Moorthy 2008). This is, at least in part, achieved through the binding of these molecules to a transcription factor known as the aryl hydrocarbon receptor (AHR) (Nebert and Russell 2002). After binding, the ligand-activated cytoplasmic AHR translocates to the nucleus where, after forming a heterodimer with another nuclear factor, interacts with AHR response elements (AHREs; also called XREs (xenobiotic response elements)) located upstream of *CYP1A* genes in multiple copies (Corchero et al. 2001). This cascade of reactions results in the initiation of transcription. CYP1A enzymes are expressed in many tissues, with CYP1A2 being expressed principally in the liver and CYP1A1 mainly in extrahepatic tissues including lungs and intestine (Moorthy 2008).

Members of the CYP1A family have major roles in the biotransformation of a variety of xenobiotics (Daly 2003). They are responsible for both detoxification and metabolic activation of PAHs and aromatic and heterocyclic amines found in cigarette smoke, combustion products and charcoal grilled food (Nebert and Russell 2002; Nebert et al. 2004).

Furthermore, CYP1A enzymes have also been studied extensively based on their potential influence in the development of chemically induced cancers in humans (Urban et al. 2001). This is based on their roles in the metabolic activation of substrates into electrophilic reactive intermediates, leading to the activation of procarcinogens and induction of toxicities (Ioannides and Parke 1990; Moorthy 2008). Therefore, analysing the genetic variation underlying interindividual differences in regulation and metabolic activity of CYP1A enzymes is clinically very important.

5.1.5. CYP1A1

CYP1A1 is a well-studied cytochrome P450 enzyme, which is present in all vertebrates (Nelson 1996). CYP1A1 has very low constitutive expression. However, induced expression by PAHs occurs in virtually all tissues and cell types (Nebert et al. 2004). Among all different tissues, lung is thought to be the main place where it is expressed (Chang and Waxman 2006). Furthermore, many of the inducers of CYP1A1 are in turn metabolised by the very same enzyme (Nebert et al. 2004).

CYP1A1 is not best known for pharmaceutical drug metabolism even though it can metabolise some drugs (Moorthy 2008). Except for the metabolism of eicosanoids, which are endogenous signalling molecules (Nebert and Karp 2008), CYP1A1 is known to be mostly involved in the metabolism of various PAH procarcinogens such as benzo[a]pyrene (Beresford 1993; Perera 1997).

Among all xenobiotic metabolising enzymes in CYP 1-3 families, which are mostly functionally polymorphic (Ingelman-Sundberg 2004a, 2004b), an exception is the CYP1A1 enzyme, which, along with CYP2E1, is evolutionary well conserved. The reason for this conservation is thought to be due to the importance of these enzymes in endobiotic metabolism (Ingelman-Sundberg 2002).

The cDNA of *CYP1A1* was cloned (2.6 kb) by Jaiswal et al. (1985) and the peptide sequence of the enzyme was 512 amino acids long. The structure of the gene was later revealed by Kajawiri et al. (1986). This gene spans a region of 6kb and has seven exons, where the first exon is untranslated (5'UTR).

Analyses of sequence variation within *CYP1A1* have been undertaken (www.cypalleles.ki.se). However, most studies resulted in the identification of polymorphisms that were either rare or private. In addition, due to the potential significance of this gene in tumourigenesis, many studies have concentrated on the relationship of *CYP1A1* polymorphisms and various cancers (especially lung cancer). For example, a *CYP1A1* variant (rs4646903; previously known as MspI) at the 3' end of the gene, which is associated with elevated expression, has been shown to be more frequent in patients with lung cancer (Xu et al. 1996). It is worth mentioning that some studies have also refuted such an association (Moorthy 2008).

Recently, systematic studies of *CYP1A1* gene polymorphisms were undertaken by Solus et al. (2004), Jiang et al. (2005) and Jorge-Nebert et al. (2009). Solus et al. (2004) resequenced all exons and flanking intronic regions in 11 major drug metabolising and carcinogen-activating *CYP* genes including *CYP1A1* and *CYP1A2* in 93 samples, which included individuals of Caucasian, African and Asian descent. Out of a total of 19 polymorphisms detected, 17 were in transcribed regions with eight being non-synonymous and one being synonymous. The other eight were found in the 3' UTR (four were novel), while no polymorphisms were found in the 5' UTR. Interestingly, although non-coding DNA (intronic and intergenic) regions usually tolerate more variation, only two polymorphisms were found in such regions. Furthermore, each SNP had an overall minor allele frequency (MAF) of less than 5%, except for one SNP (MAF = 0.11), which was only found in non-Africans. Jiang et al. (2005) carried out a similar study but only found four SNPs (non-synonymous) in the coding region, all being identified by Solus et al. (2004), along with eight 3' UTR SNPs. The lack of identification of the other five exonic SNPs reported by Solus et al. (2004) in this study is most likely due to the low sample size used (n = 24). In none of the studies was a common genetic variant *CYP1A1* with serious deficiency of enzyme activity identified.

Since sub-Saharan African individuals have not been well represented in these studies, by resequencing 750 samples (twelve population groups) distributed across sub-Saharan Africa, I investigated whether sub-Saharan African populations were near fixation for the ancestral variant of *CYP1A1* or have frequent variants (> 5%) which were not identified in previous studies due to relatively small sample sizes. This study will be of great importance as it has the potential to discover many more polymorphisms in this gene, which might explain the interindividual phenotypic variation observed.

5.1.6. CYP1A2

CYP1A2 is another CYP enzyme which has been studied extensively. It is constitutively expressed in the liver (Guengerich et al. 1999) and also induced by PAHs in liver and other tissues such as the gastrointestinal tract, pancreas and brain (Jorge-Nebert et al. 2010). CYP1A2 is involved in the metabolism of many environmental compounds such as heterocyclic amines (Nebert et al. 2004). Furthermore, most carcinogenic arylamines are also known to be metabolised by this enzyme (Butler et al. 1989; Guengerich et al. 1992).

CYP1A2 is also responsible for the oxidative metabolism of many drugs, as well as caffeine (Butler et al. 1989), e.g., theophylline (Sarkar et al. 1992), paracetamol (Patten et al. 1993) and lidocaine (Orlando et al. 2004). To date, many endogenous substrates have also been identified for the CYP1A2 enzyme such as bilirubin (Zaccaro et al. 2001), melatonin (Facciola et al. 2001) and uroporphyrinogen (Moorthy 2008). The constitutive expression of CYP1A2 in mammals further suggests that this enzyme serves some important endogenous function (Urban et al. 2001).

The cDNA of *CYP1A2* was cloned (3.1 kb) by Jaiswal et al. (1986) and the peptide sequence of the enzyme was found to be 516 amino acids long including the initiation methionine. The structure of this 7.8 kb-long gene was later identified by Ikeya et al. (1989). *CYP1A2*, which is the duplicated copy of *CYP1A1*, has also seven exons, with the first exon being untranslated (5'UTR). The level of conservation among these two genes is even more evident since exons 2, 4, 5 and 6 are very similar in both nucleotide composition and total number of bases (Ikeya et al. 1989).

There are around 60-fold differences in hepatic CYP1A2 activity at the interindividual level (Butler et al. 1989; Eaton et al. 1995; Nebert et al. 2004). Although polymorphisms have been reported to alter CYP1A2 enzyme activity in both directions (www.cypalleles.ki.se), most polymorphisms were either genotyped in a single ethnic group (Nakajima et al. 1999; Sachse et al. 1999) or found to be absent in the wider population (Allorge et al. 2003). Therefore, none of the polymorphisms reported could unequivocally explain the remarkable differences in levels of CYP1A2 enzyme activity.

However, with the aid of new technology, many studies have taken the resequencing approach to screen the whole gene for novel polymorphisms in relatively larger samples comprised of different ethnic groups (Solus et al. 2004, Jiang et al. 2005 and Jorge-Nebert et al. 2009). Even though it has been established that the people of sub-Saharan Africa have the highest human genetic diversity (Tishkoff et al. 2009), the African contribution to the samples in these studies has started to increase only in the recent years (Browning 2009).

The *CYP1A2* gene has recently been resequenced by the Environmental Genome Project (EGP) in a set of sub-Saharan African samples (Yoruba, n=12) and 11 SNPs were identified (http://egp.gs.washington.edu/). In the present study, four of the SNPs characterised in EGP were typed to generate functional haplotypes extending over most of the gene (Intron1 to 3'UTR) in samples from West Africa and the areas covered by the expansion of Bantu-speaking peoples (EBSP).

5.1.7. Aims

1) To characterise variation in frequency of *CYP1A2* haplotypes among peoples of Africa, using a four-SNP-defined haplotype hypothesised to capture all or almost all variation in the coding and flanking intronic regions of the gene.

2) To resequence the exons and flanking introns of *CYP1A1* in the EBSP Ascertainment Plate and the Ethiopian Ascertainment Plate (see Methods below), to characterise variation in the coding and flanking intronic regions of this gene in Africa.

3) To assess the predicted phenotypic effect(s) of variation identified in 1) and 2) and to draw inferences concerning drug efficacy and safety having regard to healthcare policies in Africa.

4) To compare *CYP1A1* and *CYP1A2* data produced in this study with those published in the literature and publically available databases (see Methods below). To compare: a) EBSP with Ethiopia and with other human data, b) *CYP1A1* and *CYP1A2* and c) data from humans and other primates.

5) To assess evidence of selective pressure by applying tests of departure from neutrality.

6) Compare the pattern of variation among African ethnic groups (EBSP and Ethiopian ascertainment plates) for *CYP1A1* and *CYP1A2* with each other and with the patterns for other genes and genetic systems (e.g., NRY and mtDNA) to establish similarity or dissimilarity of patterns.

5.2. Materials and Methods

5.2.1. Samples

5.2.1.1. Sub-Saharan Africa

An ‘ascertainment plate’ (a panel of DNA samples from seven population groups or locations: Sena (Mozambique; n = 51), Mambela (Somie, Cameroon; n = 65), Asante and Bulsa (Ghana; n = 57), Shewa Arabs (Lake Chad; n = 65), Chewa (Malawi; n = 50), Bakongo (Brazzaville, Republic of the Congo; n = 55) and Sudanese Kordofanians; n = 30) was assembled for use in genotyping and resequencing. These populations collectively are distributed throughout West Africa and the regions of the EBSP plus Sudan. DNA concentrations of a large set of samples from all populations were assessed by genomic DNA gel electrophoresis and samples that contained the greatest quantity of DNA were selected from the above groups. The DNA concentration of each sample was adjusted to approximately 1ng/µl by dilution with dH_20 and comparison with DNA standards of known concentration. A similar ascertainment panel was also assembled from five major ethnic groups of Ethiopia (Afar (n = 76), Amhara (n = 77), Anuak (n = 76), Maale (n = 76) and Oromo (n = 76)), representing a North East-South West transect (For further details see Browning 2009).

All buccal swabs were collected anonymously with informed consent. Sociological data were also collected from each individual including age, birthplace, self-declared cultural identity, first language, second language and in most cases religion, as well as similar information on the individual’s father, mother, paternal grandfather and maternal grandmother. The samples were classified into groups primarily by cultural identity, first language spoken, then by place of collection. Where collections from a particular group were made in more than one location, locations are represented by averages of coordinates.

DNA from Congolese samples was extracted using the Gentra protein precipitation method (Gentra Systems, Minneapolis). Previously collected buccal swab DNA samples from West Africa and areas covered by the EBSP were extracted by standard phenol-chloroform method. For details of the Gentra method, see Appendix A.

5.2.1.2. Environmental Genome Project

The Environmental Genome Project (EGP) of the NIEHS at Washington University (http://egp.gs.washington.edu/) has resequenced over one hundred genes involved in metabolism, of which 17 are CYPs including *CYP1A1* and *CYP1A2*. These two genes have been resequenced in a total of 95 samples from the following ethnic affiliations: Nigerian (Yoruba) (n =12), African American (n =15), European (Utah residents, USA) (n= 22), Hispanic (n=22), Japanese (n=12) and Chinese (Han) (n=12). The resequencing data was retrieved from the EGP database and included in the analyses of this study.

5.2.2. Polymerase Chain Reaction (PCR) and Sequencing of *CYP1A1* Exons

5.2.2.1. PCR of *CYP1A1* Exons

CYP1A1 exons 3, 4, 5 and 6 and their flanking introns were amplified in two PCR fragments. This was achievable since these exons are both relatively small (range of 81-110 bp) and in close proximity to each other (see Table 5.1 for primer details). Each fragment was amplified separately in 10-µl reaction volumes containing 1 µl (~ 1 ng) of template DNA, 1.5 ng of each primer (forward and reverse), 1.6 µl (50 µM) dNTPs, 9.3 nM TaqStart monoclonal antibody (BD Biosciences Clontech, Oxford, UK), 1 µl of 10x Taq buffer and 0.13 units of Taq DNA polymerase (HT Biotech, Cambridge, UK). All samples (96-well plates) were then placed on a thermocycler under the following conditions: denaturation at 95°C for 5 min, followed by 35 cycles of 95°C for 45 s, 66°C for 45 s and 72°C for 45 s. The final step of the PCR program was a 7-min extension at 72°C before a 30-min hold at 4°C.

Table 5.1. Primer details for the PCR fragments amplifying exons 3-6 of *CYP1A1*

PCR fragment	Primer Name	Sequence (5' - 3')	Length (bp)	Tm (°C)	GC %
Exons 3 & 4 (520 bp)	1A1-Ex34-F	AAGCTGGGACAACAGCCTCAG	21	60.7°C	57.1%
	1A1-Ex34-R	ACACAGGGACAAGATGGATGCAGG	24	61.9°C	54.2%
Exons 5 & 6 (537 bp)	1A1-Ex56-F	GTAGTGGCTCCCTTCAAAGGGGTC	24	62.3°C	58.3%
	1A1-Ex56-R	CCATGGACAGGAGGATCAATGC	22	59.1°C	54.5%

5.2.2.2. PCR purification

PCR products were purified by mixing 30 µl of a 1 to 2 mix of water and HM-MC (40% PEG- 8000, 1 M NaCl, 2 mM Tris-HCl (pH 7.5), 0.2 mM EDTA, 3.5 mM $MgCl_2$) to each sample. The mixtures were then centrifuged at 2240 g for 45 min, followed by centrifuging the inverted PCR plate for 1 min at 13 g in order to discard the supernatant. 150 µl of 70 % ethanol was added to each sample and the plate was centrifuged at 2240x g for 25 min. The resultant supernatant was discarded by the same method as above. Samples were then dried on a thermocycler at 65 C for 5 min, followed by eluting each sample with 30 µl of distilled water. Finally, the eluted samples were mixed gently to resuspend the DNA pellet.

5.2.2.3. Dideoxy Sequencing of *CYP1A1* Exons

Sequencing was performed using the same forward and reverse primers used in the PCR reactions. PCR fragments were sequenced in 15-µl reaction volumes containing 6 µl of the purified PCR product, 0.16 µM of sequencing primer, 0.75 µl of BigDye termination mix v1.1 (Applied Biosystems) and 5 µl of the accompanied sequencing buffer. The samples were placed on a thermocycler under the following conditions: 25 cycles of 96°C for 10 seconds, 55°C for 5 seconds and 60°C for 4 minutes.

5.2.2.4. Post-Sequencing Treatment

To purify the sequencing reaction products, 2.5µl of 125 mM EDTA and 30 µl of 100% ethanol were added to each sample, thoroughly mixed and left at room temperature for 10 min. The mixture was centrifuged for 60 min at 2240x g, followed by centrifuging the inverted PCR plate at 13x g for 1 min to discard the supernatant. 30 µl of 70 % ethanol was then added to each sample and centrifuged for 10 minutes at 2240x g. The resultant supernatant was again discarded by inverting the PCR plate and centrifuging for 1 min at 13x g. Samples were then dried for 5 minutes at 65ºC before adding 10 µl of Hi-Di formamide (Applied Biosystems) prior to electrophoresis. Finally, samples were run on an ABI 3730 genetic analyser and analysed using Sequencher 4.7 software (Gene Codes Corporation, USA).

5.2.2.5. Analysis of Output

The chromatogram of each sample was visually inspected across its whole length of sequence to check for high levels of background noise that would make it difficult to call bases. Subsequently, if no significant trace of noise was found, a contig was then formed for each set of 96 samples (i.e., one sequencing plate) and each mutation site was examined by eye to determine whether it was a substitution, insertion or deletion. All samples with any ambiguous sites were resequenced. Polymorphisms which were found in the forward sequence of a sample were confirmed by resequencing the reverse strand.

5.2.3. Multiplex PCR-Minisequencing (Single Base Extension Assay)

In order to generate genotypic data for *CYP1A2* SNPs, a multiplex Polymerase Chain Reaction-Minisequencing (PCR-M) assay was developed. This method reduces the time and cost of genotyping by combining multiple PCR and minisequencing reactions in one tube.

5.2.3.1. Multiplex PCR

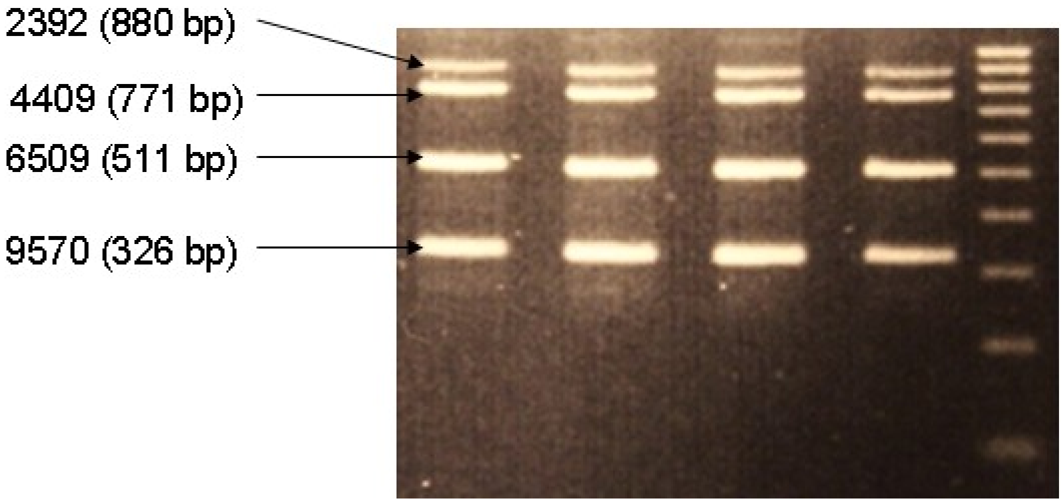
Four PCR fragments were amplified simultaneously in a single reaction using four sets of primers (see Table 5.2). All eight primers were designed using the FastPCR software (Kalendar, Lee & Schulman 2009). The 'multiplex PCR' option in this software allows for primers to be chosen with minimal molecular hybridisation within and between all oligonucleotides. The specificity of each primer was checked locally using an alignment of genic sequence between *CYP1A2* and *CYP1A1*. The primers were also checked against the whole genome for specificity, using the BLAST program (ncbi.nlm.nih.gov/blast).

Table 5.2. Primer details of the four PCR fragments in the multiplex PCR assay

PCR fragment	Primer Name	Sequence (5' - 3')	Length (bp)	Tm (°C)	GC %
2392 (880 bp)	1A2-2392-F	ACAACCCTGCCAATCTCAAGCACC	24	62.7°C	54.2%
	1A2-2392-R	CCATCTGTACCAACTGCAGGGA	22	59.8°C	54.5%
4409 (771 bp)	1A2-4409-F	CTAGCAGAGTCCTGCAATGTG	21	56.8°C	52.4%
	1A2-4409-R	AGCTGCTGGAATTATAGGGGCCTA	24	60.2°C	50.0%
6509 (511 bp)	1A2-6509-F	CCGTGAGTACATACCCCTCACGAA	24	60.6°C	54.2%
	1A2-6509-R	CATCACCTGTAACAAACGTCTTGG	24	57.3°C	45.8%
9570 (326 bp)	1A2-9570-F	TAGTAGAGACGGGTTTCACCA	21	55.6°C	47.6%
	1A2-9570-R	AAGCCAGTCACAAAAGACCACTC	23	58.4°C	47.8%

A 20-µl PCR reaction mix was comprised of the following reagents: 2µl (~ 1µM) of genomic DNA (gDNA) and 0.06 µl (50 µM) of forward and reverse primers for each fragment, in the presence of 2 µl of 10x Taq buffer, 5 µl (25 µM) of dNTPs, 23.3 nM TaqStart monoclonal antibody (BD Biosciences Clontech, Oxford, UK) and 0.33 units of Taq DNA polymerase (HT Biotech, Cambridge, UK). The reaction mixtures were incubated at 95ºC for 4 min, followed by 33 cycles of 95ºC for 1 min, 63.5ºC for 1 min and 72ºC for 1 min. The final step of the PCR was a 7-min extension at 72ºC. The PCR products were then analyzed electrophoretically on an ethidium bromide-stained 2% Agarose gel (see Figure 5.3).

Figure 5.3. Scan of an agarose gel showing the result of the multiplex PCR.



Note: The molecular DNA marker has 10 bands of sizes 1000 to 100 bp from top to bottom.

5.2.3.2. Multiplex PCR Purification

Excess primers and dNTPs were removed from multiplex PCR reactions using Exonuclease I (Exo I) and Shrimp Alkaline Phosphatase (SAP). Exo I catalyses the cleavage of single-stranded oligonucleotides (i.e., primers) into their constituent nucleotides in a 3' → 5' fashion, while SAP removes the 5'- phosphate group of dNTP molecules. The degradation of PCR primers and de-phosphorylation of unincorporated dNTPs enables the subsequent minisequencing reaction to be implemented in an efficient manner with almost no interference.

Due to the high viscosity of the enzyme solutions and minute volumes required, a master mix was made up of SAP and Exo I with a 5 to 3 ratio respectively (in units). The purification mix per sample comprised of 0.8 units of Exo I and 1.33 units of SAP. This was added to 4 µl of PCR product and the following program was run in a thermocycler: incubation at 37ºC for 60 min, inactivation of enzymes at 75ºC for 15 min and cooling at 4ºC for 30 min. When the next step (minisequencing) was not carried out immediately, samples were stored at – 20ºC.

5.2.3.3. Minisequencing

The minisequencing reaction, which is a single-base primer extension method (Sokolov 1990; Pastinen et al. 1997), was also multiplexed and all four SNPs were typed simultaneously. Minisequencing primers were designed manually so that extended primers differed in size by a minimum of 3-4 base pairs, for better differentiation of fluorescent peaks and easier detection of alleles (see Table 5.3). A master mix of ABI Prism SNaPshot Multiplex Kit (Applied Biosystems) and primer mix (four minisequencing primers) was made in a 2:1 ratio respectively. 2.33 µl of the master mix was then added to 1 µl of purified PCR product. The thermal cycling was performed under the following conditions for 25 cycles: 96ºC for 10 s, 50ºC for 5 s and 60ºC for 30 s.

5.2.3.4. Minisequencing Purification

Extended minisequencing primers were purified using the Calf Intestinal Phosphatase (CIP). CIP removes the 5'- phosphoryl group from unincorporated fluorescent ddNTPs and prevents them from co-migrating with extended primers during electrophoresis. This purification step was performed in the following way: 1 µl of a master mix consisting of CIP, 10x CIP buffer and dH_2O (0.43 µl, 0.33 µl and 0.23 µl per sample) was added to each minisequencing reaction sample and incubated at 37ºC for 1 hour, followed by 75ºC for 15 min.

Table 5.3. Details of the primers used in the multiplex minisequencing reaction

Primer Name	Sequence (5' - 3')	Length (bp)	Orientation
EXT-2392-F	GATAAAGAATGCCCTGGGGAGG	22	Sense
EXT-4409-F	GCACAGCAAGAAGGGGCCTAGAGCCAG	27	Sense
EXT-6509-F	GAAGGCTGTTTGTCCCTGCTAGGAACTGTTTA	32	Sense
EXT-9570-R	GGGTGATGAAAATGTTCTAAAATTAATTGTGGGCCGG	37	Anti-sense

5.2.3.5. Electrophoresis of Minisequencing Products

To each cleaned-up minisequencing sample (~ 4 µl), 16 µl of Hi-Di formamide was added. The samples were then denatured at 95ºC for 4 min before they were snap-cooled on ice. Finally, samples were run on *ABI PRISM 3100 Genetic Analyser* (Applied Biosystems) and the output was analysed using *GeneScan ver 3.7*.

5.2.4. Statistical Analysis

Deviation from Hardy-Weinberg expectations was assessed at each locus and its significance was examined using a permutation test. Pairwise linkage disequilibrium (LD) was also assessed between loci by estimating D′ and r^2. Haplotypes were inferred by parsimony and checked using the Excoffier-Laval-Balding (ELB) and Expectation Maximisation (EM) algorithms (Excoffier et al. 2003 and Excoffier & Slatkin 1995 respectively). Genetic diversity, *h,* and its standard error were estimated from unbiased formulae of Nei (1987). Mean number of pairwise nucleotide differences were also calculated for *CYP1A1* sequences in each group. Population genetic structure was estimated using Hierarchical Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992), which takes into account the evolutionary relationship between pairs of haplotypes and generates a statistic called Fixation Index (FST), when a simple structure of populations within a single group was defined. When a hierarchical structure was defined, three fixation indices F_{ST} (among all groups), F_{SC} (among groups within a geographic region) and F_{CT} (among geographic regions) were generated. Patterns of genetic differentiation were quantified using the following genetic distance measures estimated from AMOVA-based φst values (Excoffier et al. 1992; Michalakis & Excoffier 1996): a) F_{ST} (Reynolds, Weir & Cockerham 1983) (based on *CYP1A1* and *CYP1A2* haplotypes) and b) Kimura's two-parameter (K2P) model with gamma distribution of value 0.47 (Kimura 1980) (based on *CYP1A1* sequences). Significance of Fixation Indices and genetic distances were assessed by permutation test, where a null distribution is formed by calculating all possible values under rearrangements of haplotypes in the observed samples. The permutation test is usually repeated 10,000 times to give rise to the null distribution. Exact Test of Population Differentiation (ETPD) is an analogue of a Fisher's Exact test but the size of the contingency table is extended to the number of populations being compared (two in a pairwise population comparison or greater in a global test) by the total number of different haplotypes present (Raymond & Rousset 1995). Several tests of departure from neutrality such as Tajima's D (Tajima 1989), Fu's F (Fu 1997), McDonald-Kreitman (MK) test (McDonald and Kreitman 1991) and Fu and Li's D and F tests (Fu and Li 1993) were carried out in order to identify signals of selection. All the above was performed using Arlequin software version 3.0 (Excoffier et al. 2005) except for the MK and Fu & Li's D and F tests, which were performed using the DNAsp software (www.ub.edu/dnasp). LD parameters were also estimated by GOLD (Abecasis & Cookson 2000) and compared with those obtained by Arlequin. Principal Component Analysis (PCA) and Principal Coordinates Analysis (PCO) were performed, using the 'R' environment of statistical computing (www.R-project.org) by implementing the 'princomp' and 'cmdscale' functions respectively, on pairwise F_{ST} and K2P matrices and visualised using the 'plot' function. Both plots were used to visualise relationships among groups. 2x2 Fisher's exact test and test of equality of variances (F-test) were performed in R using the 'fisher.test' and 'var.test' functions. Relationships of *CYP1A1* and *CYP1A2* haplotypes were displayed via reduced-median networks constructed within Network 4.56 (Fluxus Engineering). The Freeman-Halton 2x3 extension of the Fisher exact probability test (Freeman and Halton 1951) was performed using the online calculator available at http://faculty.vassar.edu/lowry/fisher2x3.html.

5.2.5. Haplotype tagging SNP selection criteria

The four SNPs genotyped in *CYP1A2* were chosen applying the following criteria (based on the resequencing data obtained by EGP):

Select:

1. Polymorphic sites on each side, but within one thousand base pairs, of the transcribed region of *CYP1A2* that displayed greatest heterozygosity in the EGP Yoruba dataset.
2. All predicted to affect function lying between the two polymorphisms selected in 1 (i.e., non-synonymous and frame-shift mutations) unless they were singletons.

3. Select only one polymorphism in any set in total linkage disequilibrium (keeping the functional variants).

5.2.6. Sequence alignment

Alignment of the exonic and intronic sequences of the human, chimpanzee, gorilla and macaque *CYP1A* genes was available in Ensembl (release 58; May 2010). All coding sequence (CDS) alignments (exons) between human *CYP1A1* and *CYP1A2* were performed using TranslatorX (Abascal et al. 2010). The method used by this software is to translate the CDS and perform an alignment of amino acid sequences. The alignment is then used as a template to align corresponding CDS sequences based on codons and not individual nucleotides. Since alignments provide a measure of homology between nucleotide and amino-acid sequences and that CDS identity diverges more rapidly than the amino-acid sequence (due to the degenerate nature of the genetic code), coding sequence alignments based on corresponding amino-acid alignments give a more accurate estimate of divergence.

5.2.7. Prediction of the effect of non-synonymous mutations on protein functionality

The effects of non-synonymous variants were assessed based on the extent of physiochemical differences between the ancestral and derived amino-acid residues, using the classification shown in Figure 5.4.

Furthermore, a more detailed prediction was also made by using PolyPhen (Ramensky et al. 2002), which can be publicly accessed at http://genetics.bwh.harvard.edu/pph/ . PolyPhen initially checks whether the amino-acid substitution occurs in any site with annotated functional motifs. Furthermore, it aligns the protein sequence of the gene carrying the non-synoymous variant with its orthologues and determines the level of incompatibility of the substitution at that site by calculating the difference of position-specific independent counts (PSIC) (Sunyaev et al. 1999) scores of two variants. The higher a PSIC score difference, the higher the functional impact that is predicted for a particular amino-acid substitution. At PSIC ≥ 1.5, the functional damage is considered to be significant.

Figure 5.4. Amino acid molecular structures and their classification according to physiochemical properties

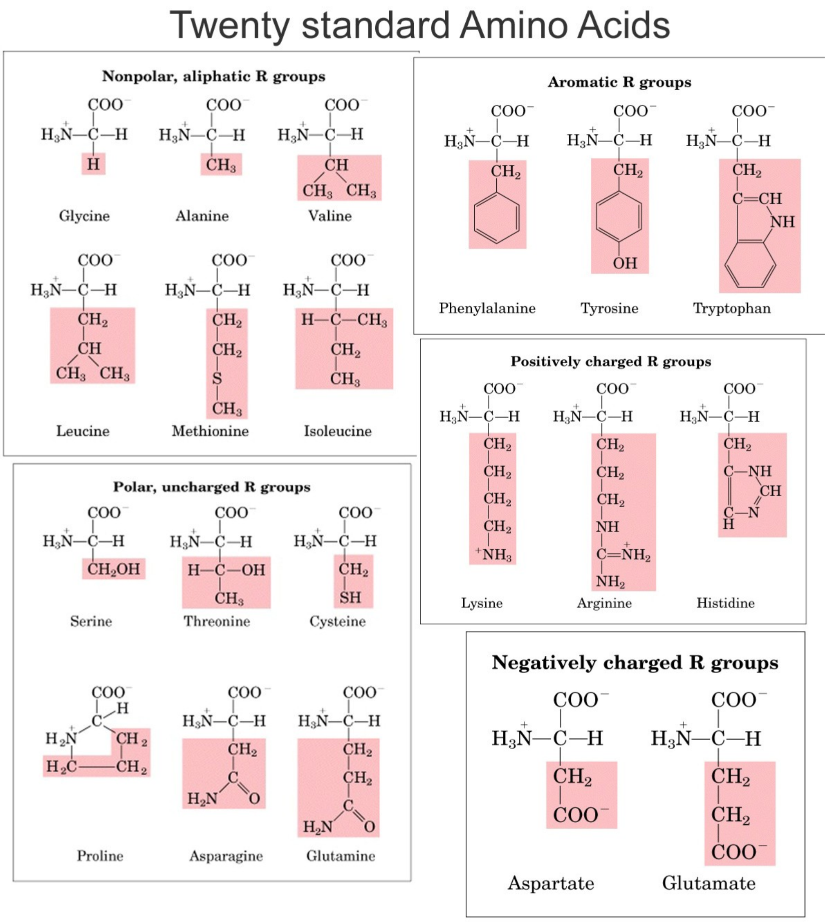


Figure obtained from http://kimwootae.com.ne.kr/apbiology/chap2.htm

5.3. Results

5.3.1. *CYP1A1*

5.3.1.1. Genetic variation in *CYP1A1*

A total of 978 bp of DNA sequence, covering exons 3 (127 bp), 4 (90 bp), 5 (124 bp) and 6 (87 bp), introns 3 (87 bp), 4 (91 bp) and 5 (145 bp) and flanking regions of introns 2 (132 bp) and 6 (94 bp), was obtained for 622 samples from twelve sub-Saharan African population groups. 14 variants were identified, of which twelve were single-base substitutions and the other two were indels. A single-base substitution was defined as a ‘Single-Nucleotide Polymorphism’ (SNP) when the minor allele was observed more than once and found at a frequency ≥ 1% in any single population group, otherwise it was defined as a rare variant. The details of all polymorphisms identified are given in Table 5.4. Exons 3 and 5 have almost the same nucleotide length, however, exon 3 carried seven (half of total) polymorphisms, while exon 5 was found to be monomorphic in 699 individuals from sub-Saharan Africa. The ratio of the number of non-synonymous to synonymous polymorphisms was 6:4. In the Ethiopian ascertainment panel, the ratio was 6:3, which was similar to that displayed in the EBSP ascertainment panel (2:1). Among the twelve single-base substitutions, nine transitions and three transversions were observed, where half of the transitions were of one type (Y; C↔T). Furthermore, 8 of 9 transition variants occurred in coding regions. The two indels identified were observed in exon 6 and only found to be present in the EBSP ascertainment panel. The 3-bp deletion at the start of the exon results in a single amino-acid (aa) residue (Threonine) deletion, while keeping the transcript in frame. However, the 1-bp insertion in the middle of the exon creates a frameshift and results in a premature stop-codon. The truncated protein (416 aa) lacks the peptide sequence encoded by exon 7 and therefore is 96 aa shorter than the ancestral transcript (512 aa). Except one (position 1390; rs34260157), all polymorphisms were only found in a heterozygote state.

Table 5.4. *CYP1A1* variants identified in this study.

Position (A in ATG is +1)	Location	Flanking sequence	Nucleotide change	Amino Acid change	dbSNP ID*
1383	Exon 3	TACCTAAGGG*CACATCCGGG	C > T		
1390	Exon 3	GGGCCACATC*GGGACATCAC	C > T	R > W	rs34260157**
1412	Exon 3	GACAGCCTGA*TGAGCACTGT	T > C	I > T	rs4987133
1418	Exon 3	CTGATTGAGC*CTGTCAGGAG	A > C	H > P	
1455	Exon 3	ACGCCAATGT*CAGCTGTCAG	C > A		
1481	Exon 3	AAGATCATTA*CATCGTCTTG	A > G	N > S	
1485	Exon 3	TCATTAACAT*GTCTTGGACC	C > T		
1545	Intron 3	GTGCTCAAGT*CCCTGACCTG	G > A		
1623	Exon 4	CTGCTATCTC*TGGAGCCTCA	C > T		
1651	Exon 4	GGTGATGAAC*CCAGGGTACA	C > T	P > S	
1988	Intron 5	ACACGGCATG*GAGACAGGGA	G > C		
2050	Exon 6	CCTAGCACAA*GAGACACAAG	CAA > -	T > -	
2091	Exon 6	CCCCAAGGGG*CGTTGTGTCT	- > G	frameshift	
2092	Exon 6	CCCCAAGGGG*GTTGTGTCTT	C > T	R > C	

* All polymorphisms except those at positions 1390 and 1412 are considered to be novel variants first identified here.
** SNP ID was obtained from the NCBI SNP database (**http://www.ncbi.nlm.nih.gov/SNP**)

As shown in Table 5.5, among the twelve single-base substitutions, only three were at polymorphic frequencies and the other nine were rare variants. Of the three SNPs, one was novel and confined to the EBSP panel and private to the Shewa Arabs in Lake Chad. The two known SNPs, which were also the most frequent of all polymorphisms, were found in both panels, with the highest frequencies observed in Ethiopian groups. Interestingly, seven rare variants were confined to the Ethiopian panel and nearly half were private to the Anuak. The most frequent SNP was observed at position 1412 (rs4987133) at 3.7% in the Anuak. Therefore, all polymorphisms identified here are not regarded to be common (i.e., they are < 5%).

Most previous studies resequencing exons 3, 4, 5 and 6 of the *CYP1A1* gene did not find any polymorphism in this region, with polymorphisms mainly confined to exons 2 (825 bp) and 7 (286 bp) (Solus et al. 2004; Jiang et al. 2005). The Environmental Genome Project (EGP) at NIEHS also found no polymorphism in this region among five diverse ethnic population groups including sub-Saharan Africans (Yoruba) and African Americans. However, two single variants (i.e., rs34260157 and rs4987133) were observed by the SNP500Cancer project (Packer et al. 2004) in a set of 24 Coriell samples of recent African origin. The predicted functional alteration of the CYP1A1 protein by the six non-synonymous SNPs identified was investigated.

Based on physiochemical properties of individual amino acids, the relative importance of each predicted change at the protein level was examined. The first SNP (R279W) is predicted to cause a dramatic change, due to a positively charged amino acid being replaced by a non-polar aromatic amino acid. The second SNP (I286T) results in the replacement of a non-polar residue with a polar one. The third (H288P) and the sixth (R405C) non-synonymous SNP result in replacement of a positively charged residue with a polar one. However, the fourth and the fifth SNP, despite changing the amino-acid residue, do not alter the physiochemical properties of their respective site on the peptide sequence.

Minor allele frequency (MAF) of each SNP was calculated in each individual population group and deviations from Hardy-Weinberg equilibrium (HWE) expectations were examined. All SNPs showed no sign of significant deviation from HWE expectations ($p > 0.05$). *CYP1A1* haplotypes were generated by visual inspection without the need to infer phase since all 1244 chromosomes were unambiguous.

Based on the haplotypes generated, estimators of linkage disequilibrium (LD) (i.e., D´ and r^2) were calculated in each individual population group. Six population groups (Sena, Ghana, Somie, Malawi, Congo and Sudan) lacked two or more polymorphic loci. The results for the rest of the groups were very similar to each other, with D´ values of 1 and r^2 values below 0.001. Furthermore, when samples from all populations were pooled together, values of D´ and r^2 were unaffected.

Nei's Gene diversity based on SNP haplotypes for the whole dataset ($n = 622$) was 0.052 ± 0.009. In individual groups, gene diversity ranged from 0.000-0.122 (mean of 0.045 and variance of 0.002) with Somie, Malawi and Sudan having no gene diversity (samples being monomorphic) and Afar having the highest value. Averaged nucleotide diversity (π) in the whole dataset was 0.005 ± 0.012 and ranged from 0.000 – 0.011 in individual groups (mean of 0.004 and variance of 0.0002) with the same order as observed for gene diversity.

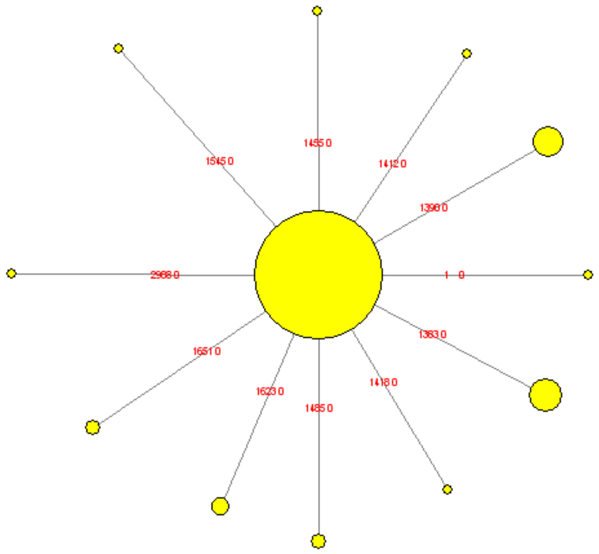
Table 5.5. Frequency distribution of variants identified in sub-Saharan African population groups.

Position (A in ATG is +1)	EBSP panel	Sena (SE)	Ghana (GH)	Somie (SO)	Malawi (MAL)	Lake Chad (LC)	Congo (CO)	Sudan (SU)	Ethiopia panel	Oromo (OR)	Amhara (AM)	Anuak (AN)	Afar (AF)	Maale (MAA)
1383	0.000* (636)	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	0.000 (104)	0.000 (58)	**0.001** **(718)**	0.000 (152)	**0.007** **(142)**	0.000 (144)	0.000 (134)	0.000 (146)
1390***	**0.004** ** **(636)**	0.000 (100)	**0.011** **(92)**	0.000 (116)	0.000 (64)	**0.020** **(102)**	0.000 (104)	0.000 (58)	**0.014** **(718)**	**0.007** **(152)**	0.000 (142)	**0.035** **(144)**	**0.030** **(134)**	0.000 (146)
1412	**0.002** **(636)**	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	**0.010** **(102)**	0.000 (104)	0.000 (58)	**0.013** **(718)**	**0.020** **(152)**	0.000 (142)	**0.007** **(144)**	**0.037** **(134)**	0.000 (146)
1418	0.000 (636)	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	0.000 (104)	0.000 (58)	**0.001** **(718)**	0.000 (152)	0.000 (142)	**0.007** **(144)**	0.000 (134)	0.000 (146)
1455	**0.002** **(636)**	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	**0.010** **(102)**	0.000 (104)	0.000 (58)	0.000 (718)	0.000 (152)	0.000 (142)	0.000 (144)	0.000 (134)	0.000 (146)
1481	0.000 (636)	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	0.000 (104)	0.000 (58)	**0.001** **(718)**	0.000 (152)	0.000 (142)	**0.007** **(144)**	0.000 (134)	0.000 (146)
1485	0.000 (636)	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	0.000 (104)	0.000 (58)	**0.001** **(718)**	0.000 (152)	0.000 (142)	**0.007** **(144)**	0.000 (134)	0.000 (146)
1545	**0.003** **(636)**	**0.010** **(100)**	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	**0.010** **(104)**	0.000 (58)	0.000 (718)	0.000 (152)	0.000 (142)	0.000 (144)	0.000 (134)	0.000 (146)
1623	0.000 (636)	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	0.000 (104)	0.000 (58)	**0.001** **(718)**	**0.007** **(152)**	0.000 (142)	0.000 (144)	0.000 (134)	0.000 (146)
1651	0.000 (636)	0.000 (100)	0.000 (92)	0.000 (116)	0.000 (64)	0.000 (102)	0.000 (104)	0.000 (58)	**0.004** **(718)**	**0.007** **(152)**	**0.007** **(142)**	0.000 (144)	0.000 (134)	**0.007** **(146)**
1988	**0.003** **(680)**	0.000 (92)	0.000 (102)	0.000 (118)	0.000 (84)	**0.016** **(122)**	0.000 (102)	0.000 (60)	0.000 (718)	0.000 (142)	0.000 (146)	0.000 (138)	0.000 (112)	0.000 (140)
2050	**0.004** **(680)**	0.000 (92)	0.000 (102)	0.000 (118)	**0.012** **(84)**	0.000 (122)	**0.020** **(102)**	0.000 (60)	0.000 (718)	0.000 (142)	0.000 (146)	0.000 (138)	0.000 (112)	0.000 (140)
2091	**0.001** **(680)**	0.000 (92)	0.000 (102)	0.000 (118)	0.000 (84)	0.000 (122)	0.000 (102)	**0.017** **(60)**	0.000 (718)	0.000 (142)	0.000 (146)	0.000 (138)	0.000 (92)	0.000 (102)
2092	0.000 (680)	0.000 (92)	0.000 (102)	0.000 (118)	0.000 (84)	0.000 (122)	0.000 (102)	0.000 (60)	**0.001** **(718)**	**0.007** **(142)**	0.000 (146)	0.000 (138)	0.000 (112)	0.000 (140)

*The number of chromosomes scanned at each polymorphic locus is shown in parenthesis.
**Variable sites in each population group are shown in bold type. ***Variable sites that can be defined as a SNP are shown in bold red.

A reduced-median network of *CYP1A1* SNP haplotypes was generated. Since all SNPs seem to have risen independently of each other (based on the observation that no compound haplotype was found), the network is a vivid example of a star-shape haplotype tree (see Figure 5.5).

Figure 5.5. Reduced-median network of *CYP1A1* haplotypes observed in sub-Saharan Africa.



5.3.1.2. Selection tests

In order to see whether the *CYP1A1* gene has been under selection, five selective neutrality tests were implemented in each population group. Fu's Fs, McDonald-Kreitman test and Fu & Li's D and F tests could not be calculated when all samples in a population group were monomorphic or polymorphisms were not present in the coding region. As shown in Table 5.6, both D and Fs are significant in three Ethiopian populations along with Shewa Arabs in Lake Chad. However, the other two Ethiopian groups show no significance for either of the two statistics. Among population groups, Sena, Ghana and Congo only show high significance for Fs and not for D. No significant departure of neutrality was found using the McDonald-Kreitman test. Fu and Li's tests were not significant in some populations but both were consistently significant in three Ethiopian populations (Amhara, Anuak and Oromo).

Table 5.6. Selective neutrality test statistics and corresponding significance.

Population group	Tajima's test	Fu's test	McDonald-Kreitman test	Fu and Li's test	
	D (p-value)	Fs (p-value)	Fisher's exact test p-value	D (p-value)	F (p-value)
Sena	-1.040 (0.130)*	**-2.130 (<0.001)****	NA	-2.00 (0.05 < p < 0.10)	-2.00 (0.05 < p < 0.10)
Ghana	-1.042 (0.118)	**-2.111 (0.029)**	1.00	-2.00 (0.05 < p < 0.10)	-1.99 (0.05 < p < 0.10)
Malawi	0.000 (1.000)	NA	NA	NA	NA
Somie	0.000 (1.000)	NA	NA	NA	NA
Lake Chad	**-1.488 (0.015)**	**-4.271 (0.001)**	1.00	-1.46 (p > 0.10)	-1.80 (p > 0.10)
Congo	-1.028 (0.121)	**-2.219 (<0.001)**	NA	-2.03 (0.05 < p < 0.10)	-2.02 (0.05 < p < 0.10)
Sudan	0.000 (1.000)	NA	NA	NA	NA
Amhara	**-1.796 (0.002)**	**-8.802 (<0.001)**	1.00	**-2.94 (p < 0.05)**	**-2.86 (p < 0.05)**
Afar	**-1.339 (0.038)**	**-4.475 (<0.001)**	0.33	0.67 (p > 0.10)	0.18 (p > 0.10)
Anuak	**-1.627 (0.008)**	**-5.938 (<0.001)**	0.40	**-2.76 (p < 0.05)**	**-2.81 (p < 0.05)**
Oromo	-0.981 (0.167)	-1.858 (0.085)	0.33	**-3.42 (p < 0.02)**	**-3.40 (p < 0.02)**
Maale	-1.000 (0.122)	**-2.428 (0.014)**	1.00	-2.09 (0.05 < p < 0.10)	-2.06 (0.05 < p < 0.10)

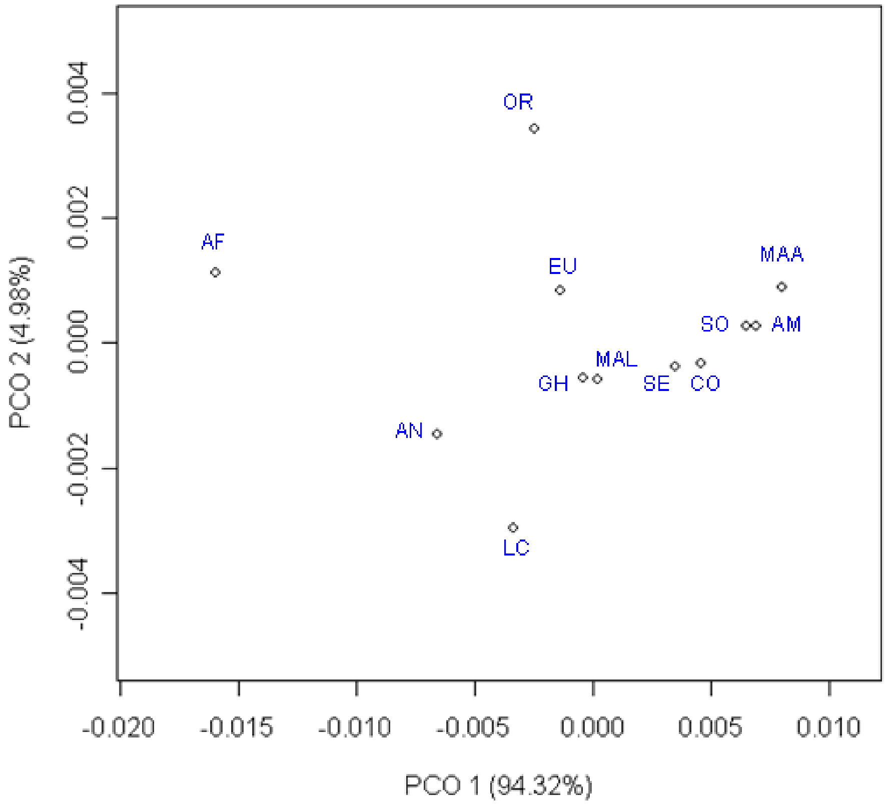
* Negative values of Tajima's D, Fu's Fs and Fu & Li's D and F tests indicate departure from neutrality (larger negative statistic values display more significance) ** Significant neutrality statistics and corresponding p-values are shown in bold type

The same tests were also performed for the whole dataset. Tajima's D and Fu's Fs were -1.929 (p < 0.001) and -3.4 x 10^{38} (p < 0.001) respectively. Fisher's exact test p-value of McDonald Kreitman test was still non-significant (p = 1). Fu and Li's selective parameters were more significant than when estimated in each population (D = -3.99 (p <<0.02) and F = -3.91 (p << 0.02)). Since most p-values obtained from different selection tests are given as intervals, the upper boundary value was used to obtain a conservative measure when carrying out Fisher's Combined Probability Test (FCPT). The test statistic for FCPT was highly significant (P < 0.001) even though multiple testing was taken into account. Therefore, the results strongly suggest that a significant departure from neutrality has occurred in the *CYP1A1* gene.

5.3.1.3. Inter-population analysis

Based on *CYP1A1* SNP haplotypes, AMOVA fixation indices indicate that almost 99.7 % of variation observed is within the individual population groups and the amount of variation between the two panels (EBSP and Ethiopia) was < 0.1% (p = 0.72). Genetic differences among individual groups were also analysed by estimating pairwise genetic distances (K2P) based on the same haplotypes.

Figure 5.6. PCO plot of K2P pairwise distances among twelve sub-Saharan African and one European population groups.



Percentages in parentheses are the amount of variation explained by each eigen vector. Abbreviations are the same as in Table 2. EU stands for the European population (EGP).

Among the 66 pairwise comparisons implemented, only two pairwise distances were significant at 5% level (Lake Chad vs Amhara and Afar vs Amhara) and the maximum pairwise genetic distance (0.021) was also observed between Afar and Amhara. After applying a Bonferroni correction, the two distances were no longer significant. A visual representation of the genetic relationship among these groups is shown in Figure 5.6.

Furthermore, pairwise genetic differences were also assessed by ETPD using all *CYP1A1* haplotypes. The pattern observed displayed a higher level of genetic differentiation among population groups than that observed by pairwise K2P values. Lake Chad was differentiated from Congo and Maale along with Amhara, while Ethiopian populations were almost all differentiated from each other except for Oromo, which displayed no significant genetic differentiation with any other

Ethiopian population. Since all non-African populations are monomorphic for this region, no pairwise comparisons could be made among themselves since gene diversity in all populations was zero. Therefore, the European population was used as the representative of populations outside of sub-Saharan Africa. Due to lack of significant variation, all sub-Saharan African populations had non-significant genetic distances (F_{ST} <0.008, p > 0.26) with the European population.

5.3.2. CYP1A2

5.3.2.1. Genetic variation in *CYP1A2*

Four of the SNPs characterised in EGP (with relative positions of 2392, 4409, 6509 and 9570) were typed to generate haplotypes extending over most of the 7.5 kb gene (Intron1 to 3'UTR) in the EBSP ascertainment panel. The second SNP at position 4409 is a non-synonymous change in exon 3, whereas the other three SNPs are intronic. All four SNPs were shown to be variable in the ascertainment panel, but not in all population groups (see Table 5.7).

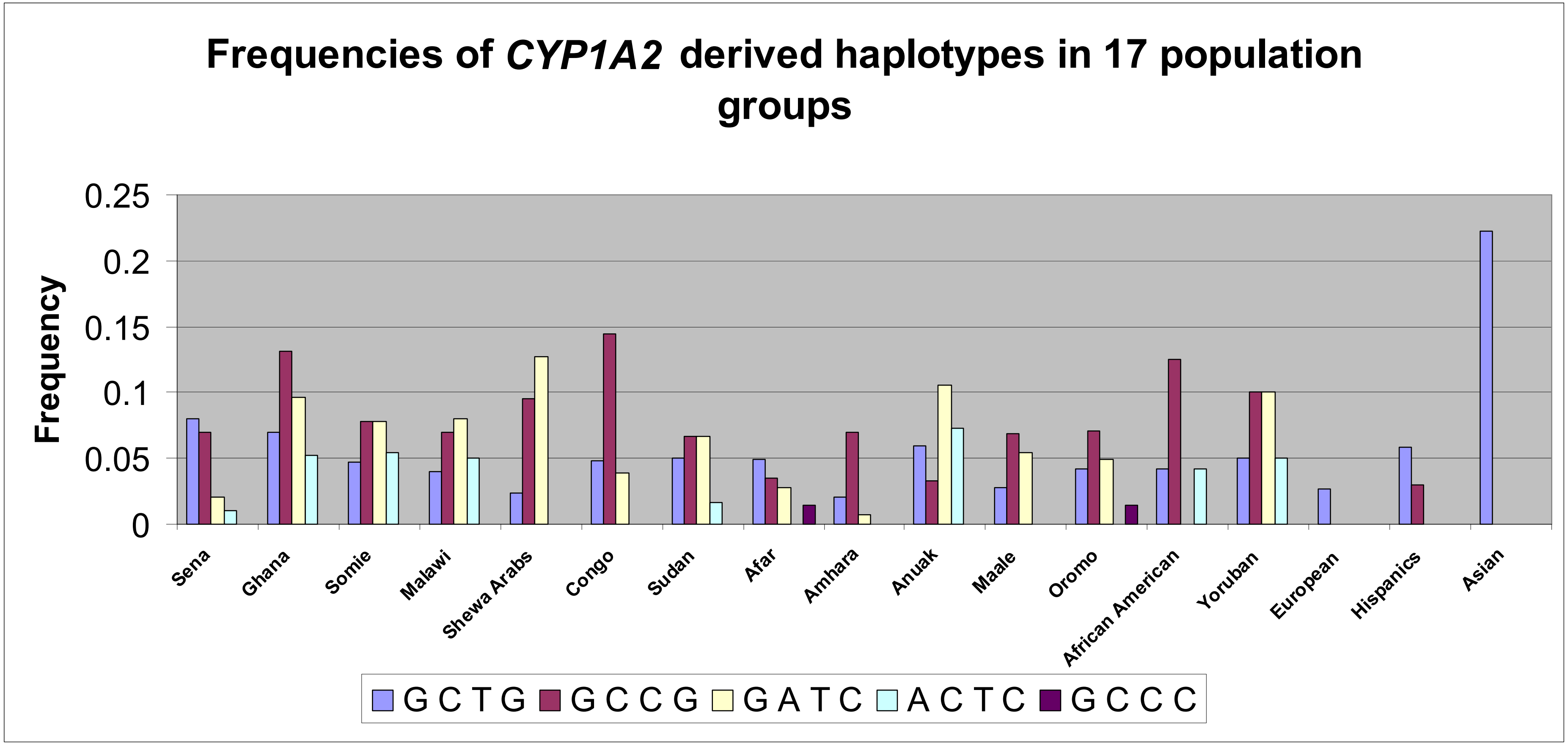
Table 5.7. Minor allele frequencies of the four SNPs in all population groups

Population group	Position on gene sequence (EGP)			
EBSP panel	2392	4409	6509	9570
Sena	**0.010**	**0.020**	**0.070**	**0.150**
Ghana	**0.053**	**0.096**	**0.132**	**0.202**
Somie	**0.055**	**0.078**	**0.078**	**0.125**
Malawi	**0.050**	**0.080**	**0.070**	**0.110**
Lake Chad	**0.000**	**0.131**	**0.092**	**0.115**
Congo	**0.000**	**0.055**	**0.136**	**0.191**
Sudan	**0.017**	**0.067**	**0.067**	**0.117**
Ethiopian panel				
Oromo	**0.000**	**0.050**	**0.086**	**0.114**
Amharic	**0.000**	**0.007**	**0.070**	**0.090**
Anuak	**0.070**	**0.105**	**0.033**	**0.092**
Maale	**0.000**	**0.055**	**0.068**	**0.096**
Afar	**0.000**	**0.028**	**0.050**	**0.085**
EGP panel				
African-Yoruba	**0.040**	**0.120**	**0.080**	**0.150**
African Americans	**0.030**	**0.030**	**0.110**	**0.190**
European	**0.000**	**0.000**	**0.000**	**0.030**
Hispanic	**0.000**	**0.000**	**0.020**	**0.080**
Asian	**0.000**	**0.000**	**0.020**	**0.220**

Figure 5.7 is a histogram of minor allele frequencies at the four SNP loci in twelve population groups. The 2392 and 9570 SNPs were found to have the lowest and highest heterozygosity in all population groups except for the Shewa Arabs in Lake Chad. The 4409 SNP was found be at the highest frequency in Lake Chad and present only in African and recent African-descent population groups (i.e. African Americans). Although the derived allele (non-synonymous change) at position 4409 (rs17861157) was present at a frequency range of 0.020-0.131 across the population groups in the ascertainment panel, this allele has not been reported in individuals without a recent African ancestry. In a parallel study (Browning 2009), allele frequencies for these four SNPs were obtained through resequencing in Ethiopian ascertainment panel. Similar minor allele frequencies were obtained to those in the EBSP region, with the exception that 2392 was present only in the Anuak (see Figure 5.7). Deviations from HWE expectations were examined for all four SNPs in each individual population group. Among all four SNPs tested in all populations, only Somie (at 4409) and Sena (at 6509) were found to show deviation from HWE expectations (p-values of 0.037 and 0.011 respectively). However, when a Bonferroni correction was applied for testing of multiple groups, no locus departed significantly from Hardy-Weinberg equilibrium in any group.

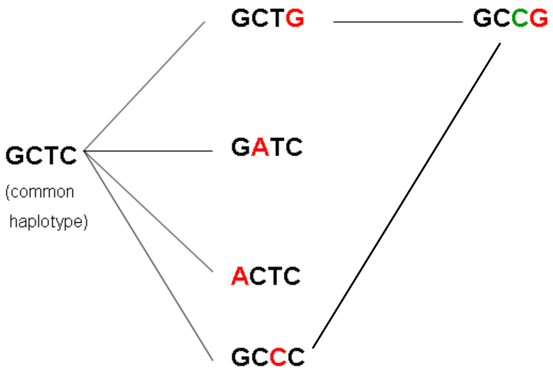
CYP1A2 haplotypes were inferred using the ELB algorithm and parsimony. Six haplotypes were identified with the common haplotype being shared with the chimpanzee and macaque (see Figure 5.8). As estimator of LD, D′ was calculated for each ascertainment panel. In the EBSP ascertainment panel, all D′ value among all loci were equal to 1, indicating strong LD across the gene. However, when D′ was calculated for the Ethiopian acertainment panel, three out of six pairwise LD were below 1 (2392-6509, 2392-9570 and 4409-6509 had D′ equal to 1)) which displays a lower level of LD in Ethiopia.

Figure 5.7. The distribution of CYP1A2 derived haplotypes in EBSP, Ethiopian and EGP population groups. The modal ancestral haplotype ranged from 0.649-0.974, where the lowest and highest frequencies were observed in the Ghanaian and European population groups (not shown).



Furthermore, haplotype GCCC was observed only in the Ethiopian ascertainment panel (Afar and Oromo each having two chromosomes). This explains the lack of LD between 6509 and 9570 in the Ethiopian ascertainment panel since all four allelic combinations between these two SNPs were observed.

Figure 5.8. A one-step mutation network of all inferred *CYP1A2* haplotypes.



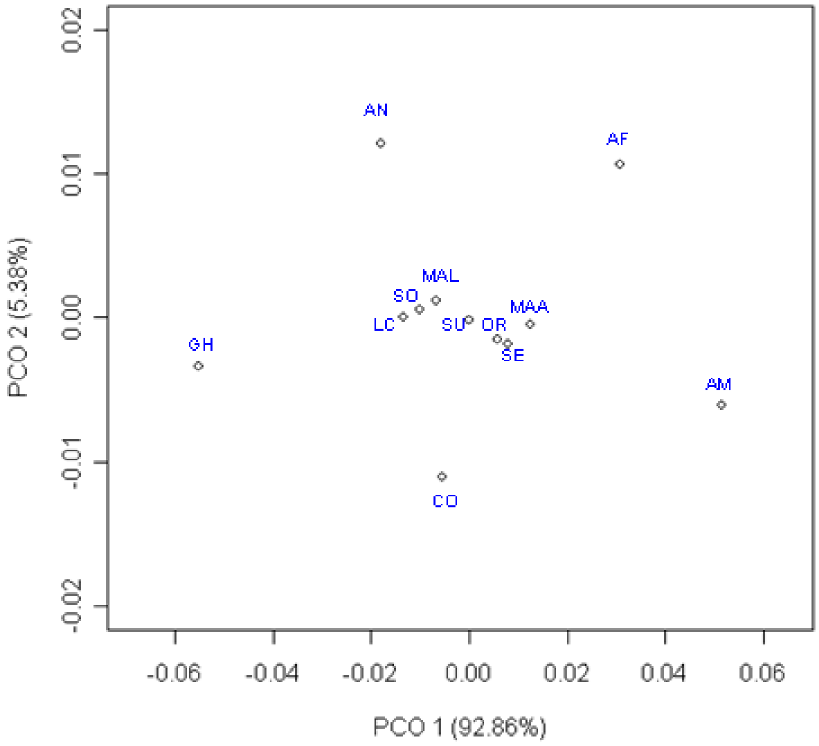
Interestingly, the non-synonymous allele (4409) was present in only one haplotype (GATC). At the protein level, the 4409 SNP (S298R) could result in a significant change in protein structure since substituting the amino acid serine for arginine could change the overall conformation, due to the increase of positive charge on the amino-acid R group.

Nei's gene diversity based on inferred *CYP1A2* haplotypes for the whole sub-Saharan African dataset (n = 727) was 0.358 ± 0.016. In individual groups, gene diversity ranged from 0.181-0.549 (mean of 0.358 and variance of 0.011), with Amhara and Ghana having the lowest and highest values respectively.

5.3.2.2. Inter-population analysis

Based on AMOVA fixation indices, although 98.3 % of variation observed was within the individual population groups, the amount of variation among population groups was highly significant ($F_{ST} = 0.017$, $p < 0.0001$). Pairwise genetic distances (F_{ST}) were estimated based on *CYP1A2* haplotypes. Among the 66 pairwise comparisons implemented, 22 pairwise distances were significant at 5% level. Among the population groups, Sudan had no significant distances, while Amhara, Afar and Anuak each had six including among themselves, except Amhara vs Afar.

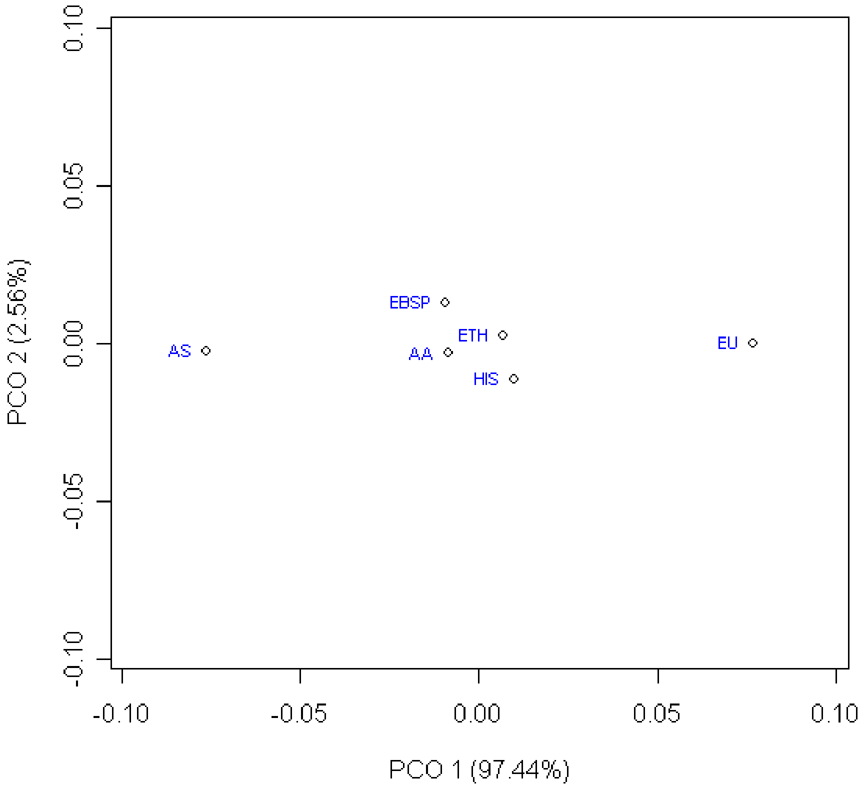
Figure 5.9. PCO plot of pairwise genetic distances (F_{ST}) of sub-Saharan African groups based on *CYP1A2* haplotypes.



Percentages in brackets are the amount of variation explained by each eigen vector. Abbreviations are the same as in Table 5.5.

The maximum pairwise genetic distance (F_{ST} = 0.105) was observed between Amhara and Ghana (see Figure 5.9). Based on the inferred haplotypes, ETPD was also used to examine pairwise genetic differentiation among groups. A similar pattern was displayed to that observed by pairwise F_{ST} values. However, rather than Amhara or Afar, Ghana and Anuak showed the highest level of differentiation from other population groups (eight significant pairwise differences ($p < 0.05$)). The sub-Saharan African populations were then pooled into the two previously defined ascertainment panels representing EBSP and Ethiopia and compared with populations outside of Africa. Based on pairwise genetic distances (F_{ST}), EBSP was significantly different from all other populations, except for the African Americans.

Figure 5.10. Visual representation of genetic distances (F_{ST}) between the EBSP and Ethiopian ascertainment panel populations and populations outside Africa based on *CYP1A2* haplotypes



AA, AS, HIS and ETH are abbreviations for African Americans, Asians, Hispanics and Ethiopians respectively.

Although significant, the genetic distance between EBSP and Ethiopia (F_{ST} = 0.012) was the lowest of all the pairwise comparisons with EBSP. Ethiopia was also found to have significant genetic distances with the European (F_{ST} = 0.030) and Asian (F_{ST} = 0.055) populations. Among all, the largest pairwise genetic distance was observed between the European and Asian populations (F_{ST} = 0.153) (see Figure 5.10).

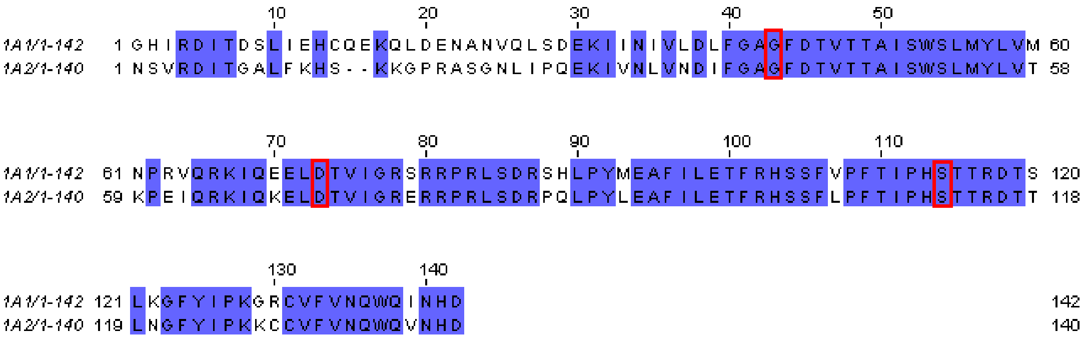
The ETPD results were similar to those obtained based on pairwise F_{ST} values, however, with a few exceptions. The pairwise genetic differences between EBSP-Hispanics and Ethiopia-European were no longer significant, while the Asians and the African Americans, which had a non-significant F_{ST} ($p = 0.161$) were genetically differentiated ($p = 0.022$).

Furthermore, when individual groups within the two ascertainment panels were compared with the European population, the average genetic distance of Ethiopian populations was 0.038 (lowest and highest observed in the Afar (0.014) and Anuak (0.085)), while EBSP populations had an average distance of 0.081 (lowest and highest observed in Sena (0.040) and Ghana (0.133)).

5.3.3. Variation in CYP1A cDNA and proteins

Using TranslatorX (Abascal et al. 2010), the cDNA sequence of exons 3, 4, 5 and 6 of both *CYP1A* genes (obtained from the NCBI CCDS database; http://www.ncbi.nlm.nih.gov/projects/CCDS/) were translated into their respective peptide sequences and aligned, based on the ClustalW alignment algorithm (see Figure 5.11).

Figure 5.11. Pairwise alignment of CYP1A protein sequence encoded by exons 3, 4, 5 and 6.



```
                       10        20        30        40        50
1A1/1-142    1 GHIRDITDSLIEHCQEKQLDENANVQLSDEKIINIVLDLFGAGFDTVTTAISWSLMYLVM  60
1A2/1-140    1 NSVRDITGALFKHS--KKGPRASGNLIPQEKIVNLVNDIFGAGFDTVTTAISWSLMYLVT  58

                       70        80        90        100       110
1A1/1-142   61 NPRVQRKIQEELDTVIGRSRRPRLSDRSHLPYMEAFILETFRHSSFVPFTIPHSTTRDTS 120
1A2/1-140   59 KPEIQRKIQKELDTVIGRERRPRLSDRPQLPYLEAFILETFRHSSFLPFTIPHSTTRDTT 118

                       130       140
1A1/1-142  121 LKGFYIPKGRCVFVNQWQINHD                                       142
1A2/1-140  119 LNGFYIPKKCCVFVNQWQVNHD                                       140
```

Identical positions between the two polypetides are highlighted in blue. First amino acid encoded by each exon is shown in a red rectangle (exon-exon boundaries).

Based on the pairwise protein alignment, exons 3 and 5 had the lowest and highest percentage identity respectively (38.1% and 87.8% respectively). Exons 4 and 6 had relatively similar percentage identities of 83.3% and 82.8% respectively. A 2x2 Fisher's exact test was used to determine whether the conservation of any of the four exons significantly differs from any other (see Table 5.8).

Table 5.8. Pairwise comparison of conservation among exons 3, 4, 5 and 6 based on percentage amino acid identities and the corresponding p values.

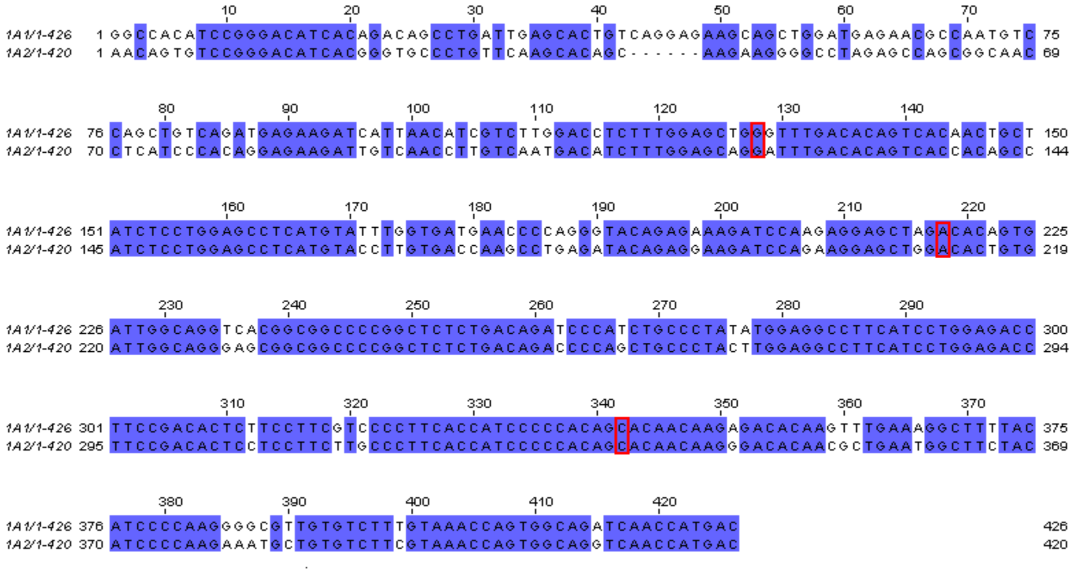
Exon 3				
Exon 4	45.2 (p = 0.0002)			
Exon 5	49.6 (p < 0.0001)	4.4 (p = 0.733)		
Exon 6	44.7 (p = 0.0002)	0.5 (p = 1)	4.9 (p = 0.731)	
	Exon 3	Exon 4	Exon 5	Exon 6

As an example, the level of conservation between *CYP1A1* and *CYP1A2* at exon 4 compared with exon 3 is (83.3 -38.1)% = 45.2% and the p-value for this difference is

0.0002. Overall, exon 3 was significantly less conserved than all other exons, even after applying the Bonferroni correction (p ≤ 0.0002).

The peptide sequence was then back translated into cDNA sequence while adhering to the same alignment output (alignment based on codons and not single nucleotides) (see Figure 5.12). Non-synonymous SNP loci identified in exons 3, 4, 5 and 6 in *CYP1A1* (this study) and *CYP1A2* (Browning 2009) were analysed in the context of the pairwise alignment of the two respective cDNA sequences covering exons 3 to 6.

Figure 5.12. Pairwise alignment of *CYP1A* cDNA sequences confined to exons 3, 4, 5 and 6.


10 20 30 40 50 60 70
1A1/1-426 1 GGCCACATCCGGGACATCACAGACAGCCTGATTGAGCACTGTCAGGAGAAGCAGCTGGATGAGAACGCCAATGTC 75
1A2/1-420 1 AACAGTGTCCGGGACATCACGGGTGCCCTGTTCAAGCACAGC------AAGAAGGGGCCTAGAGCCAGCGGCAAC 69
80 90 100 110 120 130 140
1A1/1-426 76 CAGCTGTCAGATGAGAAGATCATTAACATCGTCTTGGACCTCTTTGGAGCTGGGTTTGACACAGTCACAACTGCT 150
1A2/1-420 70 CTCATCCCACAGGAGAAGATTGTCAACCTTGTCAATGACATCTTTGGAGCAGGATTTGACACAGTCACCACAGCC 144
160 170 180 190 200 210 220
1A1/1-426 151 ATCTCCTGGAGCCTCATGTATTTGGTGATGAACCCCAGGGTACAGAGAAAGATCCAAGAGGAGCTAGACACAGTG 225
1A2/1-420 145 ATCTCCTGGAGCCTCATGTACCTTGTGACCAAGCCTGAGATACAGAGGAAGATCCAGAAGGAGCTGGACACTGTG 219
230 240 250 260 270 280 290
1A1/1-426 226 ATTGGCAGGTCACGGCGGCCCCGGCTCTCTGACAGATCCCATCTGCCCTATATGGAGGCCTTCATCCTGGAGACC 300
1A2/1-420 220 ATTGGCAGGGAGCGGCGGCCCCGGCTCTCTGACAGACCCCAGCTGCCCTACTTGGAGGCCTTCATCCTGGAGACC 294
310 320 330 340 350 360 370
1A1/1-426 301 TTCCGACACTCTTCCTTCGTCCCCTTCACCATCCCCCACAGCACAACAAGAGACACAAGTTTGAAAGGCTTTTAC 375
1A2/1-420 295 TTCCGACACTCCTCCTTCTTGCCCTTCACCATCCCCCACAGCACAACAAGGGACACAACGCTGAATGGCTTCTAC 369
380 390 400 410 420
1A1/1-426 376 ATCCCCAAGGGGCGTTGTGTCTTTGTAAACCAGTGGCAGATCAACCATGAC 426
1A2/1-420 370 ATCCCCAAGAAATGCTGTGTCTTCGTAAACCAGTGGCAGGTCAACCATGAC 420

Identical positions between the two genes are highlighted in blue. Exon boundaries (first nucleotide of the following exon) are shown with a red rectangle.

Using the Polyphen software, the position-specific independent count (PSIC) differential scores were calculated for all amino-acid substitutions and their phenotypic effects were predicted. According to the data, four non-synonymous variants resulting in amino-acid substitutions in CYP1A1 may have deleterious effects (see Table 5.9). Three of these substitutions, which have PSIC scores higher than 2.0, occur at sites conserved between the two closely related CYP1A proteins. The other substitutions, with almost no significant predicted phenotypic effect (PSIC scores below 1.5), occurred at sites that were not conserved but the PSIC score between the alleles were low with no apparent phenotypic significance. Furthermore, in only one case, at position 405 of the CYP1A1 sequence, did the derived allele result in the same amino acid (C; Cysteine) as observed in CYP1A2.

Furthermore, the cDNA sequences of both *CYP1A* genes were aligned in a pairwise manner with the genomes of three other primates (i.e., chimpanzee, gorilla and macaque), using the comparative genomic alignment tool in Ensembl. *CYP1A1* displayed 100% identity to its orthologue in chimpanzee. However, when compared with its orthologues in gorilla and macaque, it showed 99% and 94% sequence identity respectively. Human *CYP1A2* displayed 97% and 93% sequence identity when compared with gorilla and macaque. However, when compared with the chimpanzee, a lower level of conservation was observed. The chimpanzee *CYP1A2* sequence, compared with the human sequence, was reported to have a short indel (1bp) in exon 3, which results in a frameshift, causing early truncation of the protein sequence.

Table 5.9. Amino-acid substitutions predicted from non-synonymous variants in *CYP1A* exons 3 to 6 and the conservation of these residues between *CYP1A1* and *CYP1A2*.

Protein	AA change	PSIC	Phenotypic effect	AA site conservation	PSIC between CYP1A AA residues
CYP1A1	R279W	3.120	Probably damaging	YES	
	I286T	1.723	Possibly damaging	NO; I and F residues	1.013
	H288P	2.821	Probably damaging	YES	
	N309S	0.273	Benign	YES	
	P337S	2.289	Probably damaging	YES	
	R405C	0.580	Benign	NO; R and C residues	0.580
CYP1A2	S298R	0.867	Benign	NO; S and A residues	0.415
	T395M	0.489	Benign	NO; T and S residues	0.138
	N397H	0.856	Benign	NO: N and K residues	1.202

Note: At non-conserved amino-acid sites (between CYP1A proteins), alternative amino acids are shown in column five and the position-specific independent counts (PSIC) score difference between them is shown in the next column. Note all PSIC scores are below the significance threshold.

5.4. Discussion

In the present study, we performed SNP discovery and determined SNP frequencies by resequencing exons 3 to 6 of *CYP1A1*. Genetic variation in *CYP1A2* was also determined and analysed by genotyping four SNPs (spanning the entire gene) thought to define haplotypes that will capture most functional variation.

5.4.1. *CYP1A1* variation

Four exons, three introns (introns 3, 4 and 5) and two flanking introns (adjacent regions of introns 2 and 6) of *CYP1A1* were re-sequenced in seven sub-Saharan African population groups representing West Africa and the EBSP and five Ethiopian populations. Unlike previous studies, which were based on very few African samples (Solus et al. 2004; Jiang et al. 2005), this study re-sequenced approximately 1000 bp of the middle part of the *CYP1A1* gene in over 600 samples. In order to place this set of data in a worldwide context, corresponding re-sequencing data was retrieved for five populations (African Americans, Yoruba, Europeans, Hispanics and Asians) represented by the EGP panel. However, since no variation was observed in any of these populations for this segment of the *CYP1A1* gene, only one population was used for comparative analysis.

In addition to the two known polymorphisms in exon 3, identified by the SNP500 Cancer Project, another twelve variations were identified in sub-Saharan African populations. Except for two variants, found in intron 3 (rare variant) and intron 5 (SNP), all other variants occurred in coding exons. Among all 14 variations (of which three are SNPs), seven were specific to Ethiopian populations and five were specific to EBSP populations, while the two known SNPs were present in both EBSP and Ethiopian panels. The first of the two known SNPs was the most frequent among all polymorphisms identified (maximum observed frequency of 3.5%). The second SNP was also frequent in Ethiopian populations but only present as a singleton in Lake Chad. The populations carrying the highest level of variation were the Anuak and Oromo, with five polymorphisms each, along with Lake Chad, with four polymorphisms (two known SNPs in common). In contrast, Somie in Cameroon showed no variation for this segment of the *CYP1A1* gene. *CYP1A1* gene diversity and average nucleotide diversity were observed to be highest in African populations in this study, which is consistent with the observation that human genetic variation is highest in sub-Saharan Africa (Tishkoff et al. 2009). Furthermore, among sub-Saharan African populations, highest genetic diversity was observed in Ethiopian populations. This is consistent with anatomically modern human migrating out of Africa via Ethiopia and the occurrence of back-migration of Semitic speaking peoples into Ethiopia from the Near-East, which has resulted in Ethiopia becoming highly diverse genetically. The first major finding of this study is that previously reported lack of *CYP1A1* genetic variation has been due to small sample sizes and a restricted set of populations analysed.

CYP1A1 haplotypes were unambiguous in all individuals of sub-Saharan African populations since no compound heterozygote individual was observed and all variants had occurred independently on the background of the modal ancestral haplotype. LD results were also consistent with this observation since all D′ values were equal to 1 among all polymorphisms in all populations typed. In *CYP1A2*, except for two haplotypes, an identical pattern was observed where each derived allele had occurred on the background of the modal ancestral haplotype (Browning 2009).

Furthermore, the star-like network of *CYP1A1* haplotypes is consistent with the central region of the gene being under strong negative (purifying) selection. Therefore, it is likely that these exons (especially exons 4 and 5) encode parts of the peptide sequence that are essential for the metabolic function of the enzyme. This level of conservation among human populations is consistent with the proposed role of CYP1A1 in endobiotic metabolism (Ingelman-Sundberg 2002; Nebert and Karp 2008).

5.4.2. *CYP1A* sequence variation comparison

Based on the variation observed in exons 3 to 6 in *CYP1A1* (present study) and *CYP1A2* (Browning 2009) genes, the most frequent derived haplotype (after the modal ancestral haplotype) was observed at 4% in the Anuak and at 13% in the Lake Chad dataset respectively. The difference in the number of variants found in this region of *CYP1A* genes is of significance. In Ethiopia alone, six non-synonymous and three synonymous SNPs were identified in *CYP1A1* and no intronic SNPs were found. On the other hand, three non-synonymous SNPs and six intronic SNPs were identified in *CYP1A2* (Browning 2009) (Freeman-Halton test p-value = 0.007). The lack of variation in introns 3 (87 bp), 4 (91 bp) and 5 (145 bp) of *CYP1A1*, compared with *CYP1A2* (comparative sizes of 455 bp, 269 bp and 936 bp for introns 3, 4 and 5) suggests that due to their short lengths, any changes in the intronic sequence could result in splicing defects and therefore they are highly conserved. Interestingly, as in *CYP1A1*, exon 5 of *CYP1A2* carried no polymorphism (Browning 2009). This further corroborates that the amino-acids encoded by exon 5 are of significance in CYP1A2 protein structure and functionality. The frequency spectrum of variants was much wider in the *CYP1A2* gene compared with *CYP1A1*. For example, in Oromo, MAF of *CYP1A1* variants were in the range of 0.007 – 0.02, while MAF of *CYP1A2* variants ranged from 0.01 – 0.36 with MAF variances being significantly different (F = 0.0018, P < 0.0001). Overall, *CYP1A1* seems to be much less diverse than *CYP1A2*.

5.4.3. *CYP1A1* interpretation of tests for selection

Strong evidence of purifying selection is provided by the conservation of the amino acids coded for by exons 3-6 of *CYP1A1* over a long period, not only human and chimpanzee are identical, the commonality also includes the gorilla (Ensembl release 59; Aug 2010). While the tests for selection reported in the results section are consistent with such purifying selection, they lack power, as there are in all cases explanations other than purifying selection that could produce similar outputs, for example population expansion or sub-division. Evidence of conservation combined with population expansion is provided by the substantial number of polymorphic non-synonymous sites (6/428) but with none exceeding a frequency of 3.7% in any population. Paucity of data prevented evaluation of purifying selection based on relative reduction of intra-population diversity and pairwise genetic distances in coding regions with respect to non-coding regions (Hughes et al. 2003; 2005). However, based on the amino-acid sequence conservation between the closely related CYP1A proteins, it was possible to make some inferences. Peptide sequences encoded by *CYP1A* exons 4, 5 and 6 show over 80% sequence identity indicating a high level of amino-acid residue conservation between the two paralogues. Interestingly, the most conserved part of these peptide sequences was encoded by exon 5 (~ 88% identity), which was also found to be monomorphic in both genes in Ethiopia. A strong negative correlation ($r = -0.909$, $p < 0.05$) was observed between *CYP1A* exonic sequence identity and sum of exonic variants identified in each exon. This result indicates that the less conserved the coding exon is, the more variation it can tolerate. Therefore, it can be suggested that exons 4, 5 and 6 have been under strong purifying selection since they carry very few low frequency (≤ 2%) or no variants. A consequence of these observations is that it is reasonable to expect that possession of non-synonymous mutations in any of exons 4, 5, or 6 has a high chance of adverse consequences for the health of the individual, particularly if in the homozygous or compound heterozygous state. No homozygous or compound heterozygous individuals were identified.

Purifying selection at these conservative non-synonymous loci was also evidenced by the observation that the modal ancestral haplotype (frequency range of 93.3% – 100%) in all populations encodes amino-acid residues that have been conserved throughout mammalian evolution. Furthermore, PSIC scores of predicted amino-acid changes by non-synonymous variants in highly conserved regions were much higher than those in less conserved regions. Interestingly, the PSIC scores between alternative amino acids at non-conserved sites were also lower than the significance threshold, providing further evidence of conservation.

5.4.4. *CYP1A2* gene haplotypes

The frequency distribution of the four SNPs thought to cover most variation within the *CYP1A2* gene was determined. SNP 2393 (intron 1) was observed only in the Anuak among Ethiopian groups, but was widely spread among the EBSP population groups. This is most probably due to the Anuak being closely related to the Nilo-Saharan speaking peoples from which the Bantu speakers later emerged (almost half of the individuals in the population carry the modal EBSP NRY haplogroup (E1b1a)). The frequency spectrum of the non-synonymous SNP 4409 (exon 3) was also estimated. Based on the results, this allele seems to be present only in sub-Saharan Africans and individuals with recent African ancestry. This distribution is consistent with previous data characterising worldwide populations (Solus et al. 2004; Jiang et al. 2005). However, populations outside Africa have not been typed for this SNP in large sample sizes and therefore the absence of this SNP may be due to lack of sampling power.

The 4409 SNP (S298R) could result in a significant change in protein structure since substituting the amino acid serine for arginine could change the overall conformation, due to the increase of positive charge on the amino acid R group. Since this haplotype is at a reasonable frequency (maximum of 13%) in sub-Saharan Africa and its occurrence has been dated back to the origin of anatomically modern human (Browning 2009), its absence outside of populations of recent African descent could be due either to selection or drift resulting from an extreme bottleneck associated with the migration of anatomically modern human out of Africa approximately one hundred thousand years ago (Campbell and Tishkoff 2008). SNP 6509 (intron 6) was present in all population groups except the Europeans. The presence of this allele among the Hispanics may be the result of the African contribution to their gene pool during the Atlantic slave trade (see chapter 1). Unlike the first three SNPs, SNP 9570 (3' UTR) was present in all population groups including the Europeans. LD analysis indicates that stronger allelic association is present on the 3' end of the gene than the 5' end. This is evident from significant r^2 values for only the last two SNPs ($p < 0.001$) and not in any other combination, except for the first two SNPs in Sena ($p = 0.044$). According to the inferred haplotypes and the distribution of derived alleles, the data suggest that all SNPs seem to have occurred independently on the background of the ancestral haplotype but with differential success in propagation. Furthermore, haplotype analysis shows that the SNP 4409 is present only in one haplotype (GATC) and is not co-inherited with the other three SNPs. The frequency range of this haplotype was found to be wide (0.01-0.13), but limited to sub-

Saharan African populations. Therefore, this haplotype may be of practical use in anthropological research.

Among all haplotypes, GATC is the most likely to be associated with a phenotype that is different from the others. It is, therefore, a candidate for investigation in future studies of CYP1A2 catalytic activity and group-based pharamacogenetic study in sub-Saharan Africa (Jiang et al. 2005). No other common coding-sequence haplotype has yet been observed in *CYP1A2*. Since no genetic variation has been found within and upstream of the gene that can unequivocally explain the observed phenotypic variation (Jiang et al. 2006), it is reasonable to speculate that the variation at the phenotypic level is to some extent based on gene-gene interaction (e.g., interaction of CYP1A2 with other CYP enzymes) or the involvement of environmental factors. cDNA-directed recombinant CYP1A2 protein expression (cDNA carrying the 4409 SNP) and subsequent enzymatic assay would be able to determine whether this SNP can lead to an enzyme with altered function.

5.4.5. Genetic relationships of sub-Saharan African populations

CYP1A1 gene diversity was found to be low in all sub-Saharan African population groups. Therefore, since pairwise genetic distance (F_{ST}) is a function of diversity, most pairwise distances were not significant, based on *CYP1A1* haplotypes, however, some inferences could be made. The Shewa Arabs in Lake Chad were found to be very close to the Anuak and the EBSP population groups were much more clustered than the Ethiopian populations. Consistent with this observation, diversity within EBSP populations was also very much lower than the Ethiopian populations. In particular, diversity was at its lowest in Malawi, Somie and Sudan. The higher level of differentiation among Ethiopian populations compared with EBSP populations is consistent with Ethiopia being a candidate for the region of origin of anatomically modern human and the migration of Semitic speaking peoples into the area some five millennia ago. Based on *CYP1A2* haplotypes, a very similar pattern was also observed using pairwise genetic distances, where Ethiopian populations were more diverse than EBSP populations. However, when EBSP and Ethiopian panels were represented as single populations and placed in a worldwide context, they showed great similarity and, as expected, African Americans were the closest group to both panels. Although ETPD and pairwise F_{ST} results were globally similar, some minor differences were observed.

5.4.6. Conclusion

It is widely accepted that sub-Saharan Africa is the place of origin of anatomically modern human. However, in comparison with other regions in the world, especially Europe and Asia, studies investigating the distribution of potentially pharmacogenetically informative genetic variation have been sparse and limited to a few populations with small sample sizes. This study has contributed to redressing this imbalance by characterising two *CYP* genes in multiple sub-Saharan African populations with large sample sizes. Such studies are not only of benefit to the peoples of Africa, but are also of increasing importance in the planning of healthcare in the developed world, where the number of people with recent African ancestry is growing rapidly.

In conclusion, analysis of diversity in the two *CYP1A* genes in humans and their close primate relatives suggests that while there are strong forces acting to conserve the gene sequences, the diversity displayed by the peoples of sub-Saharan Africa, and Ethiopia in particular, presents an opportunity to research the detailed metabolic functions of the enzymes, which is not well understood. In addition, it should throw light on the implications for health of functional variation in the genes. It is possible that while variable levels of expression, when expression takes place, is of most interest in predicting drug efficacy and safety, the less common variation in protein structure may be more important in determining health. In pursuit of these objectives it would be useful to study metabolic functioning (*in vitro* and *in vivo*) of the variants identified in this study and also re-sequence regulatory regions. This study demonstrates the benefits of analysing re-sequencing data in important exons in multiple populations that have retained a high level of genetic diversity, within a framework that permits evaluation of both a) intra- and inter- ethnic group diversity in humans and, b) comparisons with other closely related primates, as a preliminary step prior to wider geographic surveys based on genotypes and *in vitro* and *in vivo* expression analysis. Currently, the three peripheral exons and intronic flanking regions, as well as the upstream regulatory region, of *CYP1A1* are being resequenced with the expectation, based on resequencing of *CYP1A2* in the same individuals, that these regions will be less well conserved than exons 3-6. Completion of this work will permit a more comprehensive comparison of the evolutionary histories of the two genes and assist in understanding possible causes of variation in their expression.

6. Conclusion

This chapter discusses implications of the principal results from the four projects described in this thesis. These studies have utilised previously known and novel genetic variation in populations of sub-Saharan Africa (and some populations outside) to elucidate a) origins of a community in Colombia thought to be descended from African male slaves, b) major expansion routes (and corresponding dates) of Bantu-speaking peoples, c) 'Out of Africa' migratory pattern of the truncated variant of *CASP12* and d) *CYP1A* functional variation in sub-Saharan Africa. Finally I discuss possible further research relating to the questions raised in this thesis.

6.1. Implications for the study of human history and evolution

Sub-Saharan Africa is thought to have the highest level of genetic diversity in the world (Bowcock et al. 1994; Excoffier 2002; Tishkoff et al. 2009). However, the majority of genetic studies in this region have been based on population samples of small size and with a restricted distribution. Currently available collections of DNA from sub-Saharan groups e.g. in the Coriell collection (Coriell Institute for Medical Research, Camden, NJ, USA) are not, and make no claim to be, representative of all sub-Saharan African populations. In some studies there is insufficient power due to small sample sizes with the consequence that appropriate inferences cannot be drawn from the data (for example see Berniell-Lee et al. 2009). Sometimes samples from different groups are pooled without justification even though their ethnic identities are not the same (see Hammer et al. 1997).

In genetic anthropological studies, to draw inferences about past demographic events sample sets of groups being studied should be relatively large and assigned to the groups on well described criteria relevant to the identity of each group (e.g. a shared language, culture, origin believes and social structure). The required sample size in any study will depend on the question investigated and the genetic profile of the group being studied. Even so, it should be noted that a single genetic outcome can usually be explained by multiple alternative demographic processes.

Although linguists and anthropologists have called for more fine-scale genetic studies (MacEachern 2000), very few have been published with datasets that are both large and geographically extensive (for example see Veeramah et al. 2010). In Chapters 2 and 3, I have attempted to contribute to filling this gap.

Chapter 4 illustrates how well defined populations represented by large sample sizes together with an appropriate set of markers can allow meaningful inferences to be made. The results of this study are consistent with the Out of Africa model with the highest diversity observed in sub-Saharan Africa. A similar pattern to this has also been observed for other genes for example *CD4* (Tishkoff et al. 1996) and *FMO2* (Veeramah et al. 2008b). Furthermore, this declining diversity gradient Out of Africa pattern is also evident in the distribution of variation observed in *CYP1A* drug metabolising enzyme coding genes (Chapter 5). In this case among multiple haplotypes present in sub-Saharan Africa, only one haplotype has a presence outside Africa (excluding groups with a recent African ancestry).

Close collaboration of historians, linguists, anthropologists and geneticists in inter-disciplinary projects can result in contributions which complement each other to resolve questions related to past demographic events; potentially revealing the complex processes that shaped them. The Palenque study (Chapter 2) is an example of such collaboration in which contributions from other disciplines (e.g., history and linguistics) were of significance in the design of the study. The high resolution genetic data used to analyse routes of the EBSP (Chapter 3) also illustrate the added utility of detailed genetic analysis combined with linguistic and archaeological data compared with low resolution genetic data analysed in earlier studies of the EBSP. The power of this study was primarily a consequence of dense sampling of key regions of sub-Saharan Africa and the choice of appropriate markers selected to define the tips of NRY genealogical trees. Furthermore, the patterns described in Chapters 2 and 3, applied only to the paternally inherited NRY not to the maternally inherited mtDNA. It may be that in the EBSP there was differential greater reproductive success of the dominant migrating males and absorption of indigenous females. In the Palenque study the hypothesis tested was derived from a pre-existing oral history that related only to males. There was no similar prior hypothesis that related to the female members of the community.

Although NRY and mtDNA, due to low effective population size, resulting in greater susceptibility to genetic drift, and gender specificity have proved, on many occasions, to be good choices of genetic systems for studying historical questions (see Thomas et al. 1998; 2000; 2002) autosomal markers, if analysed in large numbers, can be much more informative when undertaking gender non-specific demographic analysis. Currently, new large scale genotyping platforms such as Illumina ® Bead Arrays and Affymetrix ® SNP have been developed which makes it more feasible to initiate studies based on autosomal markers. Relatively low cost whole genome sequencing which is likely to be available in the near future will bring even more power to such studies.

6.2. Implications for the use of pharmacogenetics in healthcare

As mentioned earlier, sub-Saharan Africa has greater human genetic diversity than any of the non African continents. As a consequence analysis of DNA from its inhabitants is very likely to reveal novel variation not reported outside Africa. Results obtained for the three autosomal genes investigated in this thesis (Chapters 4 and 5) fulfil this expectation.

Analysis of multiple Ethiopian groups in pharmacogenetic studies is always likely to yield additional insights since Ethiopia is thought to be the most likely corridor through which anatomically modern human migrated 'Out of Africa'. Consequently, it may have the highest genetic diversity of any similarly sized region in the world. (In addition, variation has also resulted from back migration into Africa e.g., the migration of the Semitic speaking peoples approximately 5,000 years ago (Campbell & Tishkoff 2008)). Re-sequencing of *CYP1A1* in multiple sub-Saharan African population groups (Chapter 5) is one example of greater genetic diversity being present in Ethiopia than in any other region sampled.

Many studies have focused on typing Eurasian biased SNPs in sub-Saharan Africa, which have resulted in discrepancies of genotype-phenotype correlation studies in African populations (for example see Misimirembwa et al. 1996a). The limitation of this approach has been demonstrated by studies showing that markers strongly associated with a phenotype in European populations are not always present in sub-Saharan African populations displaying the same phenotype (Ingram et al. 2007). Although sequencing entire genomes in multiple population groups across the world is in progress within the '1000 genomes project' (http://www.1000genomes.org/), re-sequencing thousands of samples from multiple population groups in sub-Saharan Africa is currently economically unfeasible. A more practical approach for revealing pharmacogenetically useful insights into the extent of variation present in the general population is to focus on genes that code for transporter, receptor and metabolising proteins. One such class are those responsible for the metabolism of pharmaceutical drugs as well as endobiotic metabolism. In Chapter

5 I report how I used re-sequencing data from coding regions of *CYP1A2* obtained from multiple population groups with sample sizes ≥ 50 to identify high frequency variants present in sub-Saharan Africa and to use them as tagSNPs to estimate the frequency spectrum of variant haplotypes in a wider set of populations across the sub-continent. Using this method, the cost of typing is substantially reduced since the number of markers typed is limited to a smaller set of defining polymorphisms. The objective was to assess the frequencies of potential functional variants at significant frequencies in one or more sub-Saharan African populations included in the ascertainment panels representing EBSP and Ethiopia.

The continent-wide presence of a variant with predicted functional change is not conclusive evidence of significant protein change. This deduction can only be made after the variant haplotype has been shown to produce the variant enzyme or to prevent expression. If a variant haplotype is causative of a significant difference in enzymatic activity its distribution and frequency can be assessed in different groups and guide future healthcare policies (e.g., selection of drug and dose to administer). The immediate effect of such policy should be a reduction in adverse drug reactions.

Individualised therapy is unlikely to materialise in the foreseeable future in sub-Saharan Africa, thus group-based pharmacogenetics is thought to be a plausible focus for future work (Daar & Singer 2005). Furthermore, as mentioned earlier, since characterising all populations for pharmacogenetically relevant polymorphisms is impractical, knowledge of relationships among populations and genetic distances among them at other loci, combined with linguistic and anthropological data may allow the prediction of pharmacogenetic profiles of uncharacterised populations living in the same region.

6.3. Future work

Each of the four projects described in this thesis has been performed within severe constraints of limited time and available financial resources. Further research to substantiate, or refute, the reported results is therefore warranted. Below is a short summary of, in each case, further work that would be appropriate.

Chapter 2:

- Identification of one or more NRY SNPs that are present in the Yombe but not in the Chewa with a predicted date or dates of origin prior to the Atlantic slave trade. It would then be possible to genotype Palenque NRY to see if such SNP or SNPs were present. If one or more such SNPs were present that would provide additional evidence that it is more likely that there is a Yombe contribution to the paternal ancestry of the Palenque than a Chewa contribution.

Chapter 3:

- Further characterisation of E1b1a7 chromosomes to get a better understanding of the demographic processes underlying the EBSP.
- Sampling more populations from the eastern route of EBSP from, for example, Kenya and Tanzania (thought to have E1b1a chromosomes) to fill the gap of Central-East Africa in this study.
- Typing polymorphisms in the coding region of mtDNA for better haplotype characterisation for use in identifying the maternal relationships among EBSP populations at a higher resolution.

Chapter 4:

- Re-sequence intron 4 of *CASP12* in additional well-defined sub-Saharan African populations to obtain a wider picture of truncated variant haplotypes in the region and to examine the genetic relationships among populations.
- Identification and characterisation of further STR markers in close proximity to the stop-codon locus. Since the power of discrimination increases with the number of STRs, this has the potential to give at a finer scale an indication of the distribution of truncated variant haplotypes. Thus, it may be possible to further differentiate populations (e.g. North African from Middle Eastern and European from Asian).

Chapter 5:

- Re-sequence exons 1, 2 and 7 of *CYP1A1* to obtain full length coding sequence haplotypes and to see how well these exons have been conserved.
- Type further groups outside of sub-Saharan Africa for *CYP1A2* tagSNPs. This should show whether the haplotype carrying the S298R mutation is an African specific variant or was not identified earlier in populations without a recent African ancestry due to the small sample sizes in studies in which this variant was characterised.
- cDNA-directed expression of *CYP1A2* variant haplotypes in E.coli followed by an enzymatic assay to determine differential enzymatic activity of the variant and the ancestral haplotypes.
- Re-sequence the immediate upstream (positions -1 → -1000) region of both *CYP1A* genes to identify polymorphisms in functionally annotated transcription motifs which may be associated with phenotypic variation.
- Undertaking genotype-phenotype studies. *CYP1A* haplotypes of individuals can be inferred and using well-known substrates for both enzymes the metabolic activity of each enzyme can be determined. Correlations of the two variables could contribute to assessing how effectively group-based pharmacogenetics could be applied in sub-Saharan Africa.

Appendix A: Extraction of DNA from Buccal Swabs

Gentra Protein Precipitation Method

Preparing samples for Proteinase K digestion

1.1 Prepare a sample extraction sheet of 24 samples.

1.2 Using the sample extraction sheet, order set of 24 swab samples into a plastic rack. Please note that gloves must be worn whilst handling samples.

1.3 Bleach down the workbench before starting work. Scour around the PVC tape using a scalpel blade, and ensure that the swab is detached from the lid of the swab tube. If not, carefully break the swab away from the lid.

1.4 Turn on a water bath and ensure that it is set as 56°C and that distilled water is filled in to a depth of at least 3cm.

1.5 Take an 80 µl of 10 mg/ml of proteinase K and top it up to 50ml with the Baxter sterile water into a 50ml lidded tube. Gently mix the diluted proteinase K solution.

1.6 Add 500 µl of the proteinase K solution into each of the sample tubes. The same tip can be used, as long as it does not touch the sides of the sample tubes. Ensure that all lids are securely fastened.

1.7 Place the rack into the water bath and incubate at 56°C overnight.

Preparing samples for extraction

2.1 Turn on the hood, clean the surface by bleach using the blue towels.

2.2 Get the 50xTE tube and aliquot 1ml into an orange lidded tube and top it up to 50ml with Baxter sterile water.

2.3 Turn the incubator on at 56°C and place the 1xTE tube inside it.

2.4 Place a rack in the hood and with 24 autoclaved white eppendorf tubes in it.

2.5 Aliquot 166.7 µl of protein precipitation solution (Gentra) in each eppendorf tube and close the lids loosely.

2.6 Check tubes in pair to ensure that every tube has the protein precipitation solution.

2.7 Write sample numbers on each tube using the sample extraction sheet.

2.8 Add 500 µl of digested buccal DNA (from section1) to each tube. Double check for consistency of the sample numbers on both the sample tube and the eppendorf tube and close the lids.

Extraction

3.1 Put back the digested sample tubes in a fridge.

3.2 Vortex the eppendorf tubes containing digested DNA and precipitation solution for 30 seconds.

3.3 Place tubes in a cold rack and put them in freezer for 10mins. Take out the glycogen tube with cold blue rack at the same time to thaw.

3.4 Place 24 screw top tubes in a rack.

3.5 Write samples numbers on side and top of each screw top tubes.

3.6 Add 1 µl of glycogen in each screw –top tube; you should see a drop on the side.

3.7 Take out samples from freezer and centrifuge samples for 5min with maximum speed.

3.8 Add 500 µl of 100% isopropanol to each screw top tubes. You can use the same tip for all tubes but make sure you do not go deep in tubes.

3.9 Pour off the supernatant of the centrifuged samples into their respective screw top tubes and close the lid (screw top tube). You should see a yellow/orange protein pellet at the bottom of each eppendorf tube.

3.10 Invert the screw top tubes rapidly for 15-20 seconds.

3.11 Leave the screw top tubes at room temperature for 5mins.

3.12 Centrifuge tubes for 5 mins at maximum speed.

3.13 Pour off supernatant gently; be careful to not move the DNA pellet.

3.14 Leave tubes inverted at 45 degrees on clean blue towel for 1min.

3.15 Add 500 µl of EtOH 70% to each tube (To make the 70% EtOH add 35ml of 100% EtOH and 15ml of sterile water to a 50ml tube).

3.16 Centrifuge samples for 5min at maximum speed.

3.17 Pour off supernatant from the tubes gently and make sure not to move the DNA pellet.

3.18 Leave tubes inverted at 45 degrees for 15 mins.

3.19 Elute the DNA pellet in each screw top tube with 500 µl of preheated 1xTE which you had prepared and put in the incubator.

3.20 If samples are to be used straight away, incubate the samples at 56^{o}C for about 2 hours. Otherwise put the samples in the fridge.

References

2009. *Ethnologue: Languages of the World*. Sixteenth ed. Lewis M.P. (ed). Dallas, Texas: SIL International.

Abascal F, Zardoya R, and Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38 Suppl:W7-13.

Abecasis GR and Cookson WO (2000) GOLD--graphical overview of linkage disequilibrium. *Bioinformatics.* 16 (2):182-183.

Adams SM et al. (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am.J.Hum.Genet.* 83 (6):725-736.

Allorge D et al. (2003) Identification of a novel splice-site mutation in the CYP1A2 gene. *Br.J.Clin.Pharmacol.* 56 (3):341-344.

Anderson S et al. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290 (5806):457-465.

Andrews RM et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat.Genet.* 23 (2):147.

Arnaiz-Villena A, Reguera R, and Parga-Lozano C (2009) HLA Genes in Afro-American Colombians (San Basilio de Palenque): The First Free Africans in America. *The Open Immunology Journal* 2:59-66.

Arredi B et al. (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am.J.Hum.Genet.* 75 (2):338-345.

Bandelt HJ, Forster P, and Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol.Biol.Evol.* 16 (1):37-48.

Beleza S et al. (2005) The genetic legacy of western Bantu migrations. *Hum.Genet.* 117 (4):366-375.

Bellwood P (2001) Early Agriculturalist Population Diasporas? Farming, Languages, and Genes. *Annual Review of Anthropology* 30:181-207.

Beresford AP (1993) CYP1A1: friend or foe? *Drug Metab Rev.* 25 (4):503-517.

Berniell-Lee G et al. (2009) Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol.Biol.Evol.* 26 (7):1581-1589.

Bickerton D & Escalante A (1970) Palenquero: A Spanish-based creole of northern Colombia. *Lingua,* 24, 254-267.

Bosch E et al. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am.J.Hum.Genet.* 68 (4):1019-1029.

Bowcock AM et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368 (6470):455-457.

Brehm A et al. (2002) Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. *Ann.Hum.Genet.* 66 (Pt 1):49-60.

Browning SL (2010) Human genetic variation with implications for healthcare in Ethiopian populations. Doctoral thesis, University College London (UCL).

Butler MA et al. (1989) Human cytochrome P-450PA (P-450IA2), the phenacetin O-deethylase, is primarily responsible for the hepatic 3-demethylation of caffeine and N-oxidation of carcinogenic arylamines. *Proc.Natl.Acad.Sci.U.S.A* 86 (20):7696-7700.

Buxton ILO (2006) Pharmacokinetics and Pharmacodynamics: The dynamics of drug absorption, distribution, action and elimination. In Brunton LL (ed) Goodman & Gilman's The Pharmacological Basis of Therapeutics*,* 11th ed. New York: McGraw-Hill.

Campbell MC and Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu.Rev.Genomics Hum.Genet.* 9:403-433.

Campbell MC and Tishkoff SA (2010) The evolution of human genetic and phenotypic variation in Africa. *Curr.Biol.* 20 (4):R166-R173.

Castri L et al. (2009) mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. *Am.J.Phys.Anthropol.* 140 (2):302-311.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes.* New Jersey: Princeton University Press.

Cavalli-Sforza LL et al. (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc.Natl.Acad.Sci.U.S.A* 85 (16):6002-6006.

Chang TK and Waxman DJ (2006) Enzymatic analysis of cDNA-expressed human CYP1A1, CYP1A2, and CYP1B1 with 7-ethoxyresorufin as substrate. *Methods Mol.Biol.* 320:85-90.

Coelho M et al. (2009) On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC.Evol.Biol.* 9:80.

Coon MJ (2005) Cytochrome P450: nature's most versatile biological catalyst. *Annu.Rev.Pharmacol.Toxicol.* 45:1-25.

Corchero J et al. (2001) Organization of the CYP1A cluster on human chromosome 15: implications for gene regulation. *Pharmacogenetics* 11 (1):1-6.

Cruciani F et al. (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am.J.Hum.Genet.* 70 (5):1197-1214.

Daly AK (2003) Pharmacogenetics of the major polymorphic metabolizing enzymes. *Fundam.Clin.Pharmacol.* 17 (1):27-41.

Del Castillo N (1984) El lexico negro-africano de San Basilio de Palenque [The black African lexicon of Palenque de san Basilio]. *Thesaurus,* 39, 80-169.

Denbow J (1990) Congo to Kalahari: data and hypotheses about the political economy of the western stream of the Early Iron Age. *African Archaeological Review* 8 (1):139-175.

Destro-Bisol G et al. (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol.Biol.Evol.* 21 (9):1673-1682.

Di Giacomo F et al. (2004) Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Hum.Genet.* 115 (5):357-371.

Diamond J and Bellwood P (2003) Farmers and their languages: the first expansions. *Science* 300 (5619):597-603.

Eaton DL et al. (1995) Role of cytochrome P4501A2 in chemical carcinogenesis: implications for human variability in expression and enzyme activity. *Pharmacogenetics* 5 (5):259-274.

Ehret C. 2002. *The civilisations of Africa: A history to 1800.* Viriginia: University press of Virginia.

Evans WE and McLeod HL (2003) Pharmacogenomics--drug disposition, drug targets, and side effects. *N.Engl.J.Med.* 348 (6):538-549.

Excoffier L, Pellegrini A, and Langaney A (1987) Genetics and history of sub-Saharan Africa. *American Journal of Physical Anthropology* 30:151-194.

Excoffier L, Smouse PE, and Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131 (2):479-491.

Excoffier L and Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol.Biol.Evol.* 12 (5):921-927.

Excoffier L (2002) Human demographic history: refining the recent African origin model. *Curr.Opin.Genet.Dev.* 12 (6):675-682.

Excoffier L et al. (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol.Bioinform.Online.* 1:47-50.

Facciola G et al. (2001) Cytochrome P450 isoforms involved in melatonin metabolism in human liver microsomes. *Eur.J.Clin.Pharmacol.* 56 (12):881-888.

Fischer H et al. (2002) Human caspase 12 has acquired deleterious mutations. *Biochem.Biophys.Res.Commun.* 293 (2):722-726.

Foster EA et al. (1998) Jefferson fathered slave's last child. *Nature* 396 (6706):27-28.

Freeman GH and Halton JH (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika.* 38 (1-2):141-149.

Fu YX and Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133 (3):693-709.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147 (2):915-925.

Getachew KN. 2001. *Among the Pastoral Afar in Ethiopia: Tradition, Continuity and Socio-economic Change.* Utrecht: International Books.

Goncalves R et al. (2003) Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers. *Hum.Genet.* 113 (6):467-472.

Goncalves R, Spinola H, and Brehm A (2007) Y-chromosome lineages in Sao Tome e Principe islands: evidence of European influence. *Am.J.Hum.Biol.* 19 (3):422-428.

Gonzalez FJ, Tukey RH (2006) Drug Metabolism. In Brunton LL (ed) Goodman & Gilman's The Pharmacological Basis of Therapeutics*,* 6th ed. New York: McGraw-Hill.

Granda G (1971) Sobre la procedencia africana del habla “criolla” de San Basilio de Palenque [On the African origin of speech "criolla" Palenque de san Basilio]. *Thesaurus,* 26, 84-94.

Green CF et al. (2000) Adverse drug reactions as a cause of admission to an acute medical assessment unit: a pilot study. *J.Clin.Pharm.Ther.* 25 (5):355-361.

Greenberg JH. 1963. *The languages of Africa.* Bloomington: Indiana University Publication.

Guengerich FP et al. (1992) Elucidation of catalytic specificities of human cytochrome P450 and glutathione S-transferase enzymes and relevance to molecular epidemiology. *Environ.Health Perspect.* 98:75-80.

Guengerich FP et al. (1999) Inter-individual differences in the metabolism of environmental toxicants: cytochrome P450 1A2 as a prototype. *Mutat.Res.* 428 (1-2):115-124.

Hammer MF (1995) A recent common ancestry for human Y chromosomes. *Nature* 378 (6555):376-378.

Hammer MF et al. (1997) The geographic distribution of human Y chromosome variation. *Genetics* 145:787-805.

Hammer MF et al. (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol.Biol.Evol.* 18 (7):1189-1203.

Hebert PD et al. (2003) Biological identifications through DNA barcodes. *Proc.Biol.Sci.* 270 (1512):313-321.

Heilmann LJ et al. (1988) Trout P450IA1: cDNA and deduced protein sequence, expression in liver, and evolutionary significance. *DNA* 7 (6):379-387.

Helgason A et al. (2003) A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift. *Ann.Hum.Genet.* 67 (Pt 4):281-297.

Hiernaux J. 1975. *The people of Africa.* New York: Scribner's.

Holden CJ (2002) Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc.R.Soc.Lond.B* (269):793-799.

Huang QY et al. (2002) Mutation patterns at dinucleotide microsatellite loci in humans. *Am.J.Hum.Genet.* 70 (3):625-634.

Hudson RR and Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111 (1):147-164.

Hughes AL et al. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc.Natl.Acad.Sci.U.S.A* 100 (26):15754-15757.

Hughes AL et al. (2005) Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding Loci. *Genetics* 170 (3):1181-1187.

Ikeya K et al. (1989) Human CYP1A2: sequence, gene structure, comparison with the mouse and rat orthologous gene, and differences in liver 1A2 mRNA expression. *Mol.Endocrinol.* 3 (9):1399-1408.

Ingelman-Sundberg M, Oscarson M, and McLellan RA (1999) Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *Trends Pharmacol.Sci.* 20 (8):342-349.

Ingelman-Sundberg M (2002) Polymorphism of cytochrome P450 and xenobiotic toxicity. *Toxicology* 181-182:447-452.

Ingelman-Sundberg M (2004) Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends Pharmacol.Sci.* 25 (4):193-200.

Ingelman-Sundberg M (2004) Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. *Naunyn Schmiedebergs Arch.Pharmacol.* 369 (1):89-104.

Ioannides C and Parke DV (1990) The cytochrome P450 I gene family of microsomal hemoproteins and their role in the metabolic activation of chemicals. *Drug Metab Rev.* 22 (1):1-85.

Ivanov PL et al. (1996) Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat.Genet.* 12 (4):417-420.

Jaiswal AK, Gonzalez FJ, and Nebert DW (1985) Human dioxin-inducible cytochrome P1-450: complementary DNA and amino acid sequence. *Science* 228 (4695):80-83.

Jaiswal AK, Nebert DW, and Gonzalez FJ (1986) Human P3(450): cDNA and complete amino acid sequence. *Nucleic Acids Res.* 14 (16):6773-6774.

Jiang Z et al. (2005) Toward the evaluation of function in genetic variability: characterizing human SNP frequencies and establishing BAC-transgenic mice carrying the human CYP1A1_CYP1A2 locus. *Hum.Mutat.* 25 (2):196-206.

Jiang Z et al. (2006) Search for an association between the human CYP1A2 genotype and CYP1A2 metabolic phenotype. *Pharmacogenet.Genomics* 16 (5):359-367.

Jimenez S, Martinez B, and Hernandez M. HLA antigens and gene distribution in San Basilio (SB) an isolated black population of Colombia. Human Immunology 47(1-2), 62. 1996.
Ref Type: Abstract

Jobling MA, Hurles ME, Tyler-Smith C. 2004. *Human Evolutionary Genetics: Origins, People and Disease.* Abingdon: Garland Science.

Jorge-Nebert LF et al. (2010) Analysis of human CYP1A1 and CYP1A2 genes and their shared bidirectional promoter in eight world populations. *Hum.Mutat.* 31 (1):27-40.

Kachapati K et al. (2006) Population distribution of the functional caspase-12 allele. *Hum.Mutat.* 27 (9):975.

Kaessmann H et al. (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat.Genet.* 22 (1):78-81.

Kalendar R, Lee D, Schulman AH (2009) FastPCR Software for PCR Primer and Probe Design and Repeat Search. *Genes, Genomes and Genomics* 3(1):1-14.

Karafet TM et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18 (5):830-838.

Kashuba ADM, Bertino JS (2005) Mechanisms of Drug Interactions I: Absorption, Metabolism, and Excretion. In Piscitelli SC, Rodvold KA (eds) Drug Interactions in Infectious Diseases, 2nd ed. Totowa, NJ: Humana Press.

Kawajiri K et al. (1986) Structure and drug inducibility of the human cytochrome P-450c gene. *Eur.J.Biochem.* 159 (2):219-225.

Kayser M et al. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *Int.J.Legal Med.* 110 (3):125-129.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J.Mol.Evol.* 16 (2):111-120.

King TE et al. (2007) Thomas Jefferson's Y chromosome belongs to a rare European lineage. *Am.J.Phys.Anthropol.* 132 (4):584-589.

Kivisild T et al. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am.J.Hum.Genet.* 75 (5):752-770.

Klingenberg M (2003) Pigments of rat liver microsomes. *Arch.Biochem.Biophys.* 409 (1):2-6.

Knight A et al. (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr.Biol.* 13 (6):464-473.

Lamkanfi M, Kalai M, and Vandenabeele P (2004) Caspase-12: an overview. *Cell Death.Differ.* 11 (4):365-368.

Latter BDH (1980) Genetic Differences Within and Between Populations of the Major Human Subgroups. *The American Naturalist* 116 (2):220-237.

Lazarou J, Pomeranz BH, and Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* 279 (15):1200-1205.

Luis JR et al. (2004) The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am.J.Hum.Genet.* 74 (3):532-544.

MacEachern S (2000) Genes, tribes, and African history. *Current Anthropology* 41 (3):357-384.

Manica A et al. (2007) The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448 (7151):346-348.

Marshall A (1997) Getting the right drug into the right patient. *Nat.Biotechnol.* 15 (12):1249-1252.

Masimirembwa C et al. (1996) Phenotype and genotype analysis of debrisoquine hydroxylase (CYP2D6) in a black Zimbabwean population. Reduced enzyme activity and evaluation of metabolic correlation of CYP2D6 probe drugs. *Eur.J.Clin.Pharmacol.* 51 (2):117-122.

Masimirembwa C et al. (1996) A novel mutant variant of the CYP2D6 gene (CYP2D6*17) common in a black African population: association with diminished debrisoquine hydroxylase activity. *Br.J.Clin.Pharmacol.* 42 (6):713-719.

Masimirembwa CM et al. (1993) Genetic polymorphism of cytochrome P450 CYP2D6 in Zimbabwean population. *Pharmacogenetics* 3 (6):275-280.

Maynard SJ and Smith NH (1998) Detecting recombination from gene trees. *Mol.Biol.Evol.* 15 (5):590-599.

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351 (6328):652-654.

Michalakis Y and Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142 (3):1061-1064.

Monino Y (2007) Les rôles du substrat dans les créoles et dans les langues secrètes: le cas du palenquero, créole espagnol de Colombie [The roles of substrate in the Creole and the secret languages: The case of palenquero, Creole Spanish Colombia]. In : Karl Gadelii & Anne Zribi-Hertz (eds.), *Grammaires créoles et grammaire comparative*, Saint-Denis: Presse Universitaires de Vincennes, pp. 49-72.

Moorthy B (2008) The CYP1A Subfamily. In Ioannides C (ed) Cytochromes P450: Role in the Metabolism and Toxicity of Drugs and other Xenobiotics. Cambridge: RSC Publishing.

Moran CN et al. (2004) Y chromosome haplogroups of elite Ethiopian endurance runners. *Hum.Genet.* 115 (6):492-497.

Nakagawa T et al. (2000) Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-beta. *Nature* 403 (6765):98-103.

Nakajima M et al. (1999) Genetic polymorphism in the 5'-flanking region of human CYP1A2 gene: effect on the CYP1A2 inducibility in humans. *J.Biochem.* 125 (4):803-808.

Nebert DW et al. (1987) The P450 gene superfamily: recommended nomenclature. *DNA* 6 (1):1-11.

Nebert DW and Gonzalez FJ (1987) P450 genes: structure, evolution, and regulation. *Annu.Rev.Biochem.* 56:945-993.

Nebert DW and Russell DW (2002) Clinical importance of the cytochromes P450. *Lancet* 360 (9340):1155-1162.

Nebert DW et al. (2004) Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *J.Biol.Chem.* 279 (23):23847-23850.

Nebert DW and Karp CL (2008) Endogenous functions of the aryl hydrocarbon receptor (AHR): intersection of cytochrome P450 1 (CYP1)-metabolized eicosanoids and AHR biology. *J.Biol.Chem.* 283 (52):36061-36065.

Nei M. 1987. *Molecular Evolutionary Genetics.*Columbia University Press.

Nelson DR et al. (1996) P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* 6 (1):1-42.

Nelson DR (1999) Cytochrome P450 and the individuality of species. *Arch.Biochem.Biophys.* 369 (1):1-10.

Nelson DR et al. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* 14 (1):1-18.

Nelson DR (2009) The cytochrome p450 homepage. *Hum.Genomics* 4 (1):59-65.

Ntara SJ. 1973. *The history of the Chewa.* Heintze B (ed). Wiesbaden: Franz Steiner Ver Lag GMBH.

Nurse D (2006) Bantu languages. In Brown K, Oligvie S (eds) Concise encyclopedia of languages of the world. Oxford, UK: Elsevier Ltd.

Omura T and Sato R (1962) A new cytochrome in liver microsomes. *J.Biol.Chem.* 237:1375-1376.

Omura T and Sato R (1964) The carbon monoxide-binding pigment of liver microsomes. Evidence for its hemoprotein nature. *J.Biol.Chem.* 239:2370-2378.

Orlando R et al. (2004) Cytochrome P450 1A2 is a major determinant of lidocaine metabolism in vivo: effects of liver function. *Clin.Pharmacol.Ther.* 75 (1):80-88.

Packer BR et al. (2004) SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res.* 32 (Database issue):D528-D532.

Passarino G et al. (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am.J.Hum.Genet.* 62 (2):420-434.

Pastinen T et al. (1997) Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* 7 (6):606-614.

Patten CJ et al. (1993) Cytochrome P450 enzymes involved in acetaminophen activation by rat and human liver microsomes and their kinetics. *Chem.Res.Toxicol.* 6 (4):511-518.

Pereira L et al. (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):439-458.

Pereira L et al. (2002) Bantu and European Y-lineages in Sub-Saharan Africa. *Ann.Hum.Genet.* 66 (Pt 5-6):369-378.

Perera FP (1997) Environment and cancer: who are susceptible? *Science* 278 (5340):1068-1073.

Phillips KA et al. (2001) Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. *JAMA* 286 (18):2270-2279.

Pirmohamed M and Park BK (2003) Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology* 192 (1):23-32.

Plaza S et al. (2004) Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. *Hum.Genet.* 115 (5):439-447.

Porter TD and Coon MJ (1991) Cytochrome P-450. Multiplicity of isoforms, substrates, and catalytic and regulatory mechanisms. *J.Biol.Chem.* 266 (21):13469-13472.

Prugnolle F, Manica A, and Balloux F (2005) Geography predicts neutral genetic diversity of human populations. *Curr.Biol.* 15 (5):R159-R160.

Quintana-Murci L et al. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat.Genet.* 23 (4):437-441.

Quintana-Murci L et al. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat.Genet.* 23 (4):437-441.

Ramensky V, Bork P, and Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30 (17):3894-3900.

Rang HP et al.. 2007. *Rang and Dale's Pharmacology.* 6 ed. Edinburgh: Churchill-Livingstone.

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

Reed FA and Tishkoff SA (2006) African human diversity, origins and migrations. *Curr.Opin.Genet.Dev.* 16 (6):597-605.

Reynolds J, Weir BS, and Cockerham CC (1983) Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics* 105 (3):767-779.

Richards MB et al. (1998) Phylogeography of mitochondrial DNA in western Europe. *Ann.Hum.Genet.* 62 (Pt 3):241-260.

Roewer L et al. (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci.Int.* 114 (1):31-43.

Rosa A et al. (2004) MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. *Ann.Hum.Genet.* 68 (Pt 4):340-352.

Rosa A et al. (2007) Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC.Evol.Biol.* 7:124.

Ruhlen M. 1987. *A guide to the world's languages.* Stanford, California: Stanford University Press.

Ryman N (1983) Patterns of distribution of biochemical genetic variation in salmonids: Differences between species. *Aquaculture* 33 (1-4):1-21.

Sachse C et al. (1999) Functional significance of a C-->A polymorphism in intron 1 of the cytochrome P450 CYP1A2 gene tested with caffeine. *Br.J.Clin.Pharmacol.* 47 (4):445-449.

Salas A et al. (2002) The making of the African mtDNA landscape. *Am.J.Hum.Genet.* 71 (5):1082-1111.

Salas A et al. (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am.J.Hum.Genet.* 74 (3):454-465.

Saleh M et al. (2004) Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* 429 (6987):75-79.

Saleh M et al. (2006) Enhanced bacterial clearance and sepsis resistance in caspase-12-deficient mice. *Nature* 440 (7087):1064-1068.

Sarkar MA et al. (1992) Characterization of human liver cytochromes P-450 involved in theophylline metabolism. *Drug Metab Dispos.* 20 (1):31-37.

Schwegler A (2000) The African vocabulary of Palenque (Colombia). Part 1: Introduction and corpus of previously undocumented Afro-Palenquerisms. *Journal of Pidgin and Creole Language* 15, 241-312.

Schwegler A (2002) El vocabulario africano de Palenque (Colombia). Segunda Parte: compendio de palabras (con etimologías) [The African vocabulary of Palenque (Colombia). Part Two: compendium of words (with etymologies)]. In Moñino/ Schwegler (eds). 171-227.

Schwegler A (2006) Palenquero. In Brown K, Oligvie S (eds) Concise encyclopedia of languages of the world. Oxford, UK: Elsevier Ltd.

Schwegler A (2009) Palenque(ro): the search for its African substrate. Unpublished MS. Department of Spanish and Portugese. University of California, Irvine.

Scozzari R et al. (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am.J.Hum.Genet.* 65 (3):829-846.

Scozzari R et al. (2001) Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum.Immunol.* 62 (9):871-884.

Semino O et al. (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am.J.Hum.Genet.* 70 (1):265-268.

Simoni L et al. (2000) Geographic patterns of mtDNA diversity in Europe. *Am.J.Hum.Genet.* 66 (1):262-278.

Sims LM, Garvey D, and Ballantyne J (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):457-462.

Sokolov BP (1990) Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Res.* 18 (12):3671.

Solus JF et al. (2004) Genetic variation in eleven phase I drug metabolism genes in an ethnically diverse population. *Pharmacogenomics.* 5 (7):895-931.

Sunyaev SR et al. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng* 12 (5):387-394.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123 (3):585-595.

Thomas MG et al. (1998) Origins of Old Testament priests. *Nature* 394 (6689):138-140.

Thomas MG, Bradman N, and Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum.Genet.* 105 (6):577-581.

Thomas MG et al. (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):674-686.

Thomas MG et al. (2002) Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am.J.Hum.Genet.* 70 (6):1411-1420.

Tishkoff SA et al. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271 (5254):1380-1387.

Tishkoff SA et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324 (5930):1035-1044.

Tomas G et al. (2002) The peopling of Sao Tome (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. *Hum.Biol.* 74 (3):397-411.

Torroni A et al. (2000) mtDNA haplogroups and frequency patterns in Europe. *Am.J.Hum.Genet.* 66 (3):1173-1177.

Trovoada MJ et al. (2004) Pattern of mtDNA variation in three populations from Sao Tome e Principe. *Ann.Hum.Genet.* 68 (Pt 1):40-54.

Underhill PA et al. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 7 (10):996-1005.

Underhill PA et al. (2000) Y chromosome sequence variation and the history of human populations. *Nat.Genet.* 26 (3):358-361.

Underhill PA et al. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann.Hum.Genet.* 65 (Pt 1):43-62.

Urban P et al. (2001) Cytochrome P450 (CYP) mutants and substrate-specificity alterations: segment-directed mutagenesis applied to human CYP1A1. *Biochem.Soc.Trans.* 29 (Pt 2):128-135.

Vansina J (1995) New Linguistic Evidence and 'the Bantu Expansion'. *The Journal of African History* 36 (2):173-195.

Veeramah KR et al. (2008a) Sex-Specific Genetic Data Support One of Two Alternative Versions of the Foundation of the Ruling Dynasty of the Nso' in Cameroon. *Curr.Anthropol.* 49 (4):707-714.

Veeramah KR et al. (2008b) The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa. *Pharmacogenet.Genomics* 18 (10):877-886.

Veeramah KR et al. (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.

Verdu P et al. (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr.Biol.* 19 (4):312-318.

Wade P (1995) The cultural politics of Blackness in Colombia. *American Ethnologist* 22(2):341-357.

Watson E et al. (1996) mtDNA sequence diversity in Africa. *Am.J.Hum.Genet.* 59 (2):437-444.

Weale ME et al. (2001) Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. *Hum.Genet.* 109 (6):659-674.

Weale ME et al. (2002) Y chromosome evidence for Anglo-Saxon mass migration. *Mol.Biol.Evol.* 19 (7):1008-1021.

Weale ME et al. (2003) Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography. *Genetics* 165 (1):229-234.

Weinshilboum R (2003) Inheritance and drug response. *N.Engl.J.Med.* 348 (6):529-537.

Wood ET et al. (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur.J.Hum.Genet.* 13 (7):867-876.

Xu X et al. (1996) Cytochrome P450 CYP1A1 MspI polymorphism and lung cancer susceptibility. *Cancer Epidemiol.Biomarkers Prev.* 5 (9):687-692.

Xue Y et al. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am.J.Hum.Genet.* 78 (4):659-670.

Ye S et al. (2001) An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.* 29 (17):E88.

Yeretssian G et al. (2009) Gender differences in expression of the human caspase-12 long variant determines susceptibility to Listeria monocytogenes infection. *Proc.Natl.Acad.Sci.U.S.A* 106 (22):9016-9020.

Yu N et al. (2002) Larger genetic differences within africans than between Africans and Eurasians. *Genetics* 161 (1):269-274.

Zaccaro C et al. (2001) Role of cytochrome P450 1A2 in bilirubin degradation Studies in Cyp1a2 (-/-) mutant mice. *Biochem.Pharmacol.* 61 (7):843-849.