Human Genetic Variation with Implications for Healthcare in Ethiopian Populations

Sarah Louise Browning
The Centre for Genetic Anthropology
Department of Genetics, Evolution and Environment
University College London
Ph.D.

Declaration of ownership

I, Sarah Louise Browning, 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

Cytochrome P450 *1A2* metabolizes a wide range of therapeutic drugs, including several used to treat diseases common in sub-Saharan Africa. Variation in the gene (*CYP1A2*) has been reported to be associated with differential efficacy of therapeutic drugs and adverse drug reactions. To gain a better understanding of the extent of variation in the coding and exon-flanking non-coding regions of *CYP1A2*, 762 chromosomes from members of five ethnic groups (Afar, Amhara, Anuak, Maale and Oromo) distributed in a rough north east to south west transect across Ethiopia were re-sequenced. Substantial variation was observed, much of which was novel. As a consequence, a diagnostic test based on previously known variation cannot predict functional variation in Ethiopians. Evidence of purifying selection acting on *CYP1A2* was found and coalescent date estimates of *CYP1A2* variants were old, with many pre-dating expansions of anatomically modern human out of Africa.

Variants within the transcription factor 7-like 2 gene (*TCF7L2*), which are associated with an increased risk of type 2 diabetes (T2D), were common in multiple Ethiopian populations. *TCF7L2* haplotype distribution varied among groups suggesting that T2D susceptibility may also vary, with most groups likely having a West African *TCF7L2* risk for the disease and some having more of a European *TCF7L2* risk.

Many *CYP1A2* and *TCF7L2* haplotypes can be of important predictive value in the planning and provision of healthcare. These findings are not only of benefit to native Ethiopians, but are also of increasing importance in the planning of healthcare intervention in the developed world, where growing numbers of individuals with recent Ethiopian descent are living. Comparing data with those from publicly available databases it appears that Ethiopian groups display a very high level of diversity that includes most of the common variation observed elsewhere.

Acknowledgements

I owe a special thanks to the BBSRC for funding me, and to all sample donors and individuals involved with sample collections, especially Dr Ayele Tarekegn. I would also like to thank Professor Steve Humphries, and his group, for collaborating with me on the type 2 diabetes study and for kindly giving me some samples from their type 2 diabetic Afro-Caribbean cohort.

A massive thanks to everybody who has helped me in their own little way and made my four year stint at UCL an enjoyable one. This includes (in somewhat chronological order): Krishna Veeramah, Christopher Plaster, Catherine Ingram, Claire Hodgkinson, Yuval Itan, Laura Horsfall, Andrew Loh, Ana Texieira, Lorenzo Zannette, Ian Barnes, Jutta Palmen, Lauren Holt, Larissa Kogleck, Naser Ansari Pour, Adam Powell, Lauren Johnson, Olivia Clark, Hala Elhaj, Amanthi Balasuriya, Andrea du Preez, Bryony Jones, Maha Al-Sulaimani, Farzeen Rauf, Rosemary Ekong, Olivia Creemer, Ripu Bains and Gurjeet Rajbans.

Thank you to Ros Wolfes for housing me during the early days, and Beth Codrington, Gemma Wolfes and Tom Carroll for putting up with me at 4a Digby Crescent. A big thank you to Lizzie and John Wallwork for all their kindness and generosity over the past year. Thanks to Jill for her spring cleaning, and the two pigeons which made me smile every time they waddled past the window (one of which was killed by a hawk in front of my very eyes but I won't mention that again). A very big thank you to Ossi Wallwork for his love, support, and more importantly, for letting me boot him out of his room to write up my thesis.

This sounds crazy, but thanks to all the reality TV shows, plus all the tacky programmes that I have grown to love over the past eight months or so. They have given me something to look forward to in the evenings after a hard day's writing/analysing/fixing (sad isn't it?) and allowed my brain to rest. Also, thanks to tea, coffee and biscuits!

On a more serious note, I would like to thank (in no particular order) Professor Sue Povey, Dr Nik Maniatis, Dr Mike Weale, Professor Nancy Mendell and Professor Endashaw Bekele for all their help and guidance over the past four years.

My main thanks are of course for my industrial sponsor, Dr Neil Bradman, my supervisors, Professor Mark Thomas and Professor Dallas Swallow, and my adorable parents, Angela and Mike Browning. Without their support I would simply not have been able to complete my PhD.

Contents page

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Supplementry data

List of tables

Table 2.2 and supplementary tables S1 - S12 are shown on the attached CD due to their large size.

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Supplementary data

1 Introduction

Studies in this thesis work towards characterising human genetic variation, with implications for healthcare, in Ethiopian populations. This first chapter aims to set the scene by providing a brief description of past and present day Ethiopia. The overall rationale of the thesis and an overview of each chapter are provided. Chapters 2 - 5 are self contained results chapters, each with their own set of aims and methods.

1.1 A brief history of Ethiopia

Archaeology and paleontology bring Ethiopia's history back millions of years with some of humankind's oldest traces being unearthed in Ethiopia. The most well known example is of 'Lucy', a possible ancestor of *Homo sapiens* found at Hadar in south east Ethiopia, who is thought to have lived in Ethiopia approximately 3.2 million years ago (Shreeve, 1994). Fossils of the earliest known morphological traits associated with anatomically modern human, dated to 150 – 190 thousand years ago, have also been uncovered in Ethiopia (White et al., 2003; McDougall and Fleagle, 2005; Campbell and Tishkoff, 2008).

Ethiopian dynastic history (which is reviewed in Henze, 2001; Pankhurst, 2001), according to Ethiopian legend, traditionally begins in 1000 BC with the reign of Emperor Menelik I, whose lineage is traced back to the Queen of Sheba and King Solomon. According to historical records, Ethiopia's roots date back to the Kingdom of D'mt, in the 8th century BC, which is thought to have had much contact with South Arabia. After the fall of the Kingdom of D'mt in the 4th century BC, the Kingdom of Aksum emerged as the successor state. It was an important trading nation in the north east of Africa and expanded southwards, achieving great power status by the 6th century AD. Aksum was also the first major empire to convert to Christianity. By about 1270, the Aksumites were superseded by the Zagwes, who in turn gave way to the so-called 'Solomonid Restoration'. At its height, the Ethiopian Empire, also known in Western literature as Abyssinia, included present day Northern Somalia, Djibouti, Southern Egypt, Eastern Sudan, Yemen and Western Saudi Arabia at different times in its history. During the 19th century, several migrations and conquests led to Ethiopia's borders expanding in the south, toward its modern borders. With the exception of a short-lived Italian occupation from 1936-1941, the Ethiopian Empire was the only African nation to maintain its freedom from colonial rule. In 1974, a Soviet-backed military junta (the Derg) dethroned Emperor Haile Selassie (who had ruled de facto since 1916) and founded a single party socialist state. The communist regime suffered coups, uprisings, severe drought, famine and an immense refugee problem, and was eventually overturned in 1991 by a coalition of rebel forces, the Ethiopian People's Revolutionary Democratic Front (EPRDF). An Ethiopian constitution was adopted and Ethiopia's first multiparty elections were held in 1995 (reviewed in Henze, 2001; Pankhurst, 2001). However, the predominantly Tigrayan led Ethiopian government has been marred by

allegations of ethno-centrism, increasing authoritarianism, corruption, serious human rights abuses and electoral fraud (Vestal, 1999).

1.2 Present day Ethiopia

Ethiopia (Federal Democratic Republic of Ethiopia) is a landlocked country (1,127,127 km² (www.cia.gov/library/publications/the-world-factbook/)) partly lying on the Horn of Africa in the north east of the African continent. Dijibouti and Eritrea border Ethiopia to the north, Somalia to the East, Kenya to the south and Sudan to the west (figure 1.1). The Great Rift Valley runs through the country in a rough south west to north east direction dissecting a mass of highland mountains and plateaus. Lowlands, steppes and semi-desert regions boarder the Great Rift Valley. The topography of Ethiopia contains a number of high mountain ranges (e.g. Semien Mountains and Bale Mountains) and one of the lowest African land areas (Danakil depression). Deserts are generally found along the eastern border whilst tropical forests are found to the south and extensive Afromontane forests are located in both the north and south west of the country. Ethiopia's geographical diversity leads to high variability in climate, soils, natural vegetation and patterns of population settlement.

Figure 1.1 Location and relief map of Ethiopia (derived from www.worldofjah.ning.com)

With a population estimated at 85,200,000 in July 2009, Ethiopia is the world's 15th most populous country (www.cia.gov/library/publications/the-world-factbook/). Ethiopia has a plethora of ethnic groups, each with its own cultural practices. Some ethnic groups are subdivided into smaller groups, some of which admix extensively with others while some do not, maintaining intra-marriage traditions (Freeman and Pankhurst, 2003). The country has 84 indigenous living languages, most of them Afro-Asiatic (Semitic, Cushitic, Omotic), as well as some that are Nilo-Saharan. Amhara and Oromo are Ethiopia's main languages with the former

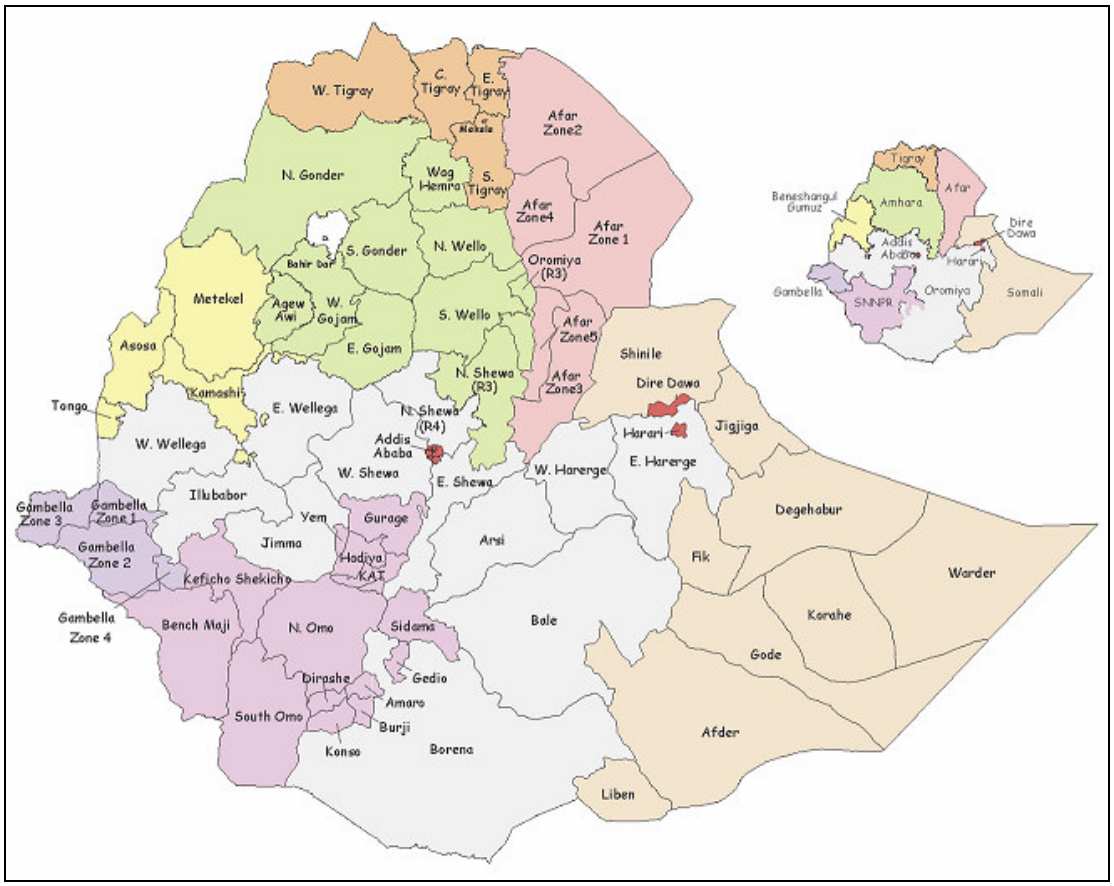
being the official language. Ethiopia's main foreign languages are English and Arabic (www.ethnologue.com). Christianity is the predominant religion but a large proportion of the population are Muslim and some still practice traditional beliefs (1994 census).

Ethiopia is divided into nine ethnically-based administrative regions which are further divided into 68 zones and two self-governing administrations, Addis Ababa (capital) and Dire Dawa (figure 1.2).

Agriculture (primarily coffee production) is Ethiopia's main economic activity and accounts for almost half of GDP, 60 % of exports, and 80 % of total employment. Frequent periods of drought and poor cultivation practices often hamper economic growth, and urban and rural poverty is widespread (www.cia.gov/library/publications/the-world-factbook/).

Poor sanitation and malnutrition lead to high incidences of communicable diseases which are worsened by a shortage of trained healthcare workers and healthcare infrastructure. There is a high risk of bacterial and protozoal diarrhea, hepatitis A and E, and typhoid fever, malaria, meningococcal meningitis, rabies and schistosomiasis. Ethiopia has a relatively low average life expectancy at birth (55 years) and infant mortality rates are relatively high with 8 % of children not surviving the first year (www.cia.gov/library/publications/the-world-factbook/).

Figure 1.2 Administrative regions (small map) and zones (large map) of Ethiopia (produced by UN emergencies unit for Ethiopia (2000))



1.3 Rationale of thesis

The Out of Africa model of human origin is a widely accepted view and postulates that modern humans (*Homo sapiens*) originated in Africa ~ 200 thousand years ago and colonised the rest of the world within the last ~ 100 thousand years (Stringer and Andrews, 1988; Tishkoff et al., 1996; Campbell and Tishkoff, 2008). Support for the Out of Africa model comes from the observation that African populations are more genetically diverse than non-African populations (Excoffier, 2002; Tishkoff and Kidd, 2004; Tishkoff et al., 2009) and that intra-population genetic diversity decreases with distance from Africa (Prugnolle et al., 2005; Handley et al., 2007; Tishkoff et al., 2009). Intra-population phenotypic diversity (measured by human skull size and shape) is also highest in African populations and is negatively correlated with distance from Africa (Manica et al., 2007; Betti et al., 2009). This pattern of variation is consistent with modern human populations experiencing serial founder effects in the course of expanding out of Africa and across the globe (Rosenberg et al., 2002; Ramachandran et al., 2005; Rosenberg et al., 2006; Jakobsson et al., 2008).

It is suggested that anatomically modern humans migrated out of Africa from the north east (possibly via Ethiopia) by crossing the Bab-el-Mandreb strait at the mouth of the Red Sea (figure 1.1) (Forster and Matsumura, 2005; Reed and Tishkoff, 2006; Campbell and Tishkoff, 2008; Tishkoff et al., 2009). Evidence of a more recent migration into Ethiopia, of Semitic speaking peoples from Arabia, is also known from genetic studies (Campbell and Tishkoff, 2008). As a consequence, it is possible that more human genetic/phenotypic variation will be observed in Ethiopians than in any other geographically contiguous indigenous groups of peoples of similar number.

Characterising the genetic distribution among ethnically and/or geographically diverse Ethiopian populations will be important for understanding the population history of Ethiopians and for assisting the reconstruction of the evolutionary history of modern humans (Campbell and Tishkoff, 2008). Given that East Africa showed the greatest level of population substructure within Africa (Tishkoff et al., 2009) and that substructure can often lead to false results (Pritchard et al., 2000; Tishkoff et al., 2009), population genetic studies in Ethiopia will also be important for the better design and interpretation of disease association and pharmacogenetic studies, not only in Ethiopians, but also in populations across the globe.

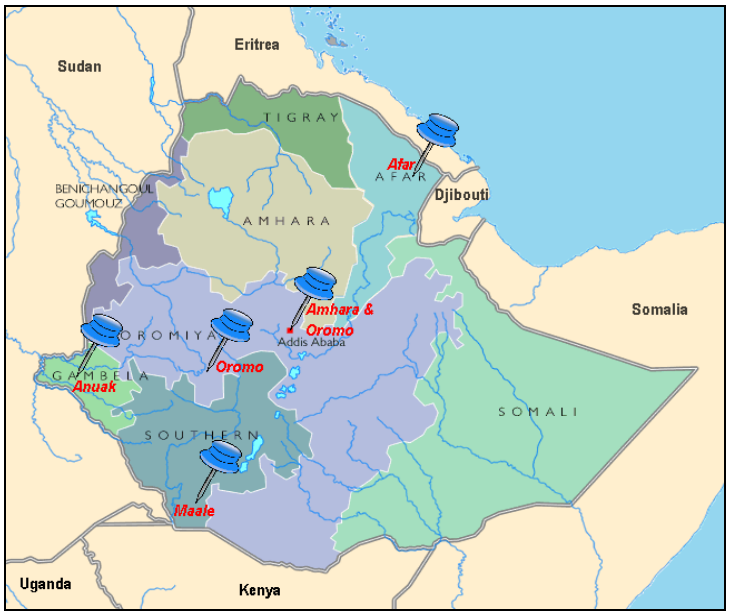
Despite the potential importance of Ethiopian population genetics, little is known about the distribution of human genetic variation among Ethiopian populations (Campbell and Tishkoff, 2008). Studies that have been undertaken have been limited to few populations and/or small sample sizes (e.g. Aklillu et al., 2003; Jiang et al., 2006; Tishkoff et al., 2009). This thesis has contributed to correcting this imbalance by analysing genetic variation in two therapeutically important genes in five Ethiopian ethnic groups (Afar, Amhara, Anuak, Maale and Oromo), collected from a roughly north east to south west transect across the country and referred to hereafter as the Ethiopian ascertainment populations.

1.4 Ethiopian ascertainment populations

Based on pairwise Fsts from Y chromosome and mtDNA hypervariable region 1 sequences, evidence suggests that the majority of genetic variation is captured by sampling a north east to south west transect across Ethiopia (unpublished data). In light of this, a total of 381 DNA samples were chosen for an 'Ethiopian ascertainment panel'. Samples were collected from five different ethnic groups inhabiting a roughly north east to south west transect across the country (figure 1.3) in order to provide a sample set suitable for an initial evaluation of diversity among the peoples of Ethiopia. A total of 76 samples were taken from each of the following ethnic groups, apart from the Amhara of which there were 77:-

1. Afar (Afar – North-East)
2. Amhara (Addis Ababa)
3. Oromo (Addis Ababa and central highlands)
4. Anuak (Gambela – West)
5. Maale (South Omo – South-West)

Figure 1.3 Collection locations of the Ethiopian ascertainment samples
(derived from http://www.younglives.org.uk/)



1.4.1 Afar

The Afar are an ethnic group from the Horn of Africa who principally reside in the eastern lowlands of the Afar Region of Ethiopia but are also found in eastern parts of Eritrea and Djibouti (Lewis, 1998). They have preserved their own rich culture and traditions and in their appearance are said to be similar to the Somali (Lewis, 1998). The total population size of the Afar has been estimated at 1,439,367 (http://www.ethnologue.com/), of which 979,367 live in

Ethiopia (1994 census). They speak the Afar language which is a member of the Cushitic branch of the Afro-Asiatic language family and is further classified as East, Saho-Afar (http://www.ethnologue.com/). Speakers also use Arabic as a second language. Alternate names for the Afar are Afaraf, Danakil, Denkel, 'Afar Af, and Adal. (Note Danakil, an Arabic term, is sometimes used by non-Afar to identify the group but is considered offensive by the Afar themselves). They are referred to as "Adal" in Amharic (http://www.ethnologue.com/). They are Muslims and traditionally nomadic herders. While the majority have remained nomadic pastoralists, raising goats, sheep, camels and cattle in the desert, some have migrated to cities and adopted urban lifestyles. The Afar are split into clan families and divided into two main classes: the Adoimara (whites) and the Asaimara (reds). The Adoimara are the dominant political class while the Asaimara are the commoners or working class. The Afar appear to be predominantly endogamous and maintain a patriarchal society (Lewis, 1998).

1.4.2 Amhara

The Amhara are an ethnic group principally residing in the Amhara region in north central Ethiopia, the central highlands of Ethiopia and Addis Ababa but are found worldwide (http://www.ethnologue.com/). They are arguably the most politically powerful group in Ethiopia and comprise an estimated 16,007,933 people, forming 30 % of the country's population (1994 census). It is thought that the Amhara migrated into the central areas of Ethiopia, including Addis Ababa, in the last millennium BC. They are believed to have both indigenous and southwest Arabian heritage and appear to be linguistically, culturally, as well as phenotypically (appearance) related to Arabs. The Amhara have undergone admixture having assimilated extensively with local populations, especially the Oromo, the Afar and the Tigray (personal communication, Ayele Tarekegn[1]). Amharic is both the language of the Amhara and Ethiopia's national language, being used in government, public media, national commerce, education and a wide variety of literature. Speakers may also use English, Arabic, Oromo or Tigrinya. Amharic is part of the Semitic branch of the Afro-Asiatic language family and is further classified as South, Ethiopian, South, Transversal, Amharic-Argobba (http://www.ethnologue.com/). A small proportion of the Amhara are Muslim but the predominant religion is Christianity, with the Ethiopian Orthodox Church playing a key role in the culture of the group and the country. A small minority of the Amhara are Protestant. Although many have moved to Addis Ababa, the vast majority of the Amhara lead a rural lifestyle and make their livelihood through arable farming, predominantly in the Ethiopian highlands. Cattle, sheep, goats, horses and donkeys are also raised (personal communication, Ayele Tarekegn).

[1] Ayele Tarekegn is an Ethiopian postdoctoral research fellow at The Centre for Genetic Anthropology in UCL. He has been conducting fieldwork in Ethiopia since 1997 and has collected buccal DNA and sociological background data from over 8000 Ethiopians from multiple ethnic groups.

1.4.3 Anuak

The Anuak are an ethnic group whose villages are scattered along river banks in the fertile lowlands in southeastern Sudan and western Ethiopia, in the Gambela Region. Members of this ethnic group number 45,665 in Ethiopia (1994 census). The Anuak speak Anuak which is part of the Nilo-Saharan language family and is further classified as Eastern Sudanic, Nilotic, Western, Luo, Northern, Anuak (http://www.ethnologue.com/). They are believed to have migrated from the area of the Upper Nile in Sudan and are culturally, linguistically, historically and religiously different from most other Ethiopians. They have traditional religious beliefs and are distinguished by their typically dark skin. The Anuak have been exposed to discrimination and marginalisation which has affected the group's access to education, healthcare and basic infrastructure. The group also remains predominantly endogamous (personal communication, Ayele Tarekegn). They are agriculturalists, growing maize and sorghum, but also earning their living through animal husbandry, fishing and hunting (http://www.ethnologue.com/).

1.4.4 Maale

The Maale are an ethnic group living in the southwest of Ethiopia in the Southern Nations, Nationalities and Peoples Region, to the northeast of Jinka town in the South Omo administrative Zone. According to the 1994 census (1994 census), they numbered 46,458. Maale is the group's language which is part of the Omotic branch of the Afro-Asiatic language family, and further classified as North, Gonga-Gimojan, Gimojan, Ometo-Gimira then Ometo (http://www.ethnologue.com/). The majority of the Maale practice their traditional religion but some are Protestant. They, like the Anuak, have typically dark skin. They are predominantly endogamous, maintaining intra-marriage traditions (personal communication, Ayele Tarekegn).

1.4.5 Oromo

The Oromo are the largest single ethnic group in Ethiopia constituting over 30 % of the country's population. According to the 1994 Ethiopian census (1994 census), members of the group numbered 17,080,318. The Oromo are found in Kenya but predominantly reside in the Oromo Region, west and central areas of Ethiopia, and along the Rift Valley escarpment east of Dessie and Woldiya. They speak the Oromo language which is part of the Cushitic branch of the Afro-Asiatic language family and further classified as East, then Oromo. Like Amharic, Oromo is widely spoken in Ethiopia. It is a trade language and is used in government, public media, national commerce, education to eighth grade and a variety of literature (http://www.ethnologue.com/). While further research is needed to fully understand the origin or origins of Oromo people, it has been postulated that they may be derived from the highlands of present-day Bale region in Southern Ethiopia (personal communication, Ayele Tarekegn). Most of the Oromo are either Muslims or Christians but some have traditional religious beliefs. The Oromo have a varied livelihood with some being agriculturalists, growing a variety of crops

including spices and coffee. Some have livestock whilst others are miners or earn their living through the tourism, textile, meat packing or refinery industries (http://www.ethnologue.com/).

1.5 Overview of thesis

Chapters 2 – 4 focus on variation observed in the Cytochrome P450 *1A2* gene (*CYP1A2*). CYP1A2 is a clinically important drug metabolising enzyme and variation in the gene has been reported to be associated with differential efficacy of therapeutic drugs and adverse drug reactions (Gunes and Dahl, 2008). The gene was chosen for study in Ethiopians because a wide range of pharmaceuticals are metabolized by CYP1A2 (Gunes and Dahl, 2008), including several used to treat diseases common in Ethiopia. Caffeine (an important component of coffee) is a well known substrate of CYP1A2 (Butler et al., 1989), and coffee was first domesticated for human use in Ethiopia (Anthony et al., 2002) and is an integral part of modern Ethiopian culture. Variation within *CYP1A2* remains largely uncharacterised within Ethiopian populations. To gain a better understanding of the extent of variation in the gene, coding and exon-flanking non-coding regions of *CYP1A2* were re-sequenced in the Ethiopian ascertainment populations.

Chapter 2 describes the variation observed in *CYP1A2* in the Ethiopian ascertainment populations. In total, 49 variable sites of different types were found, 30 of which are novel. Nine non-synonymous changes (seven of which are novel) and one synonymous change were found in the coding region. Haplotype analysis of the entire gene revealed 55 different haplotypes, only three of which were previously reported. When haplotypes were constructed using only non-synonymous polymorphisms, so as to restrict the haplotype set to those most likely to affect enzyme structure/function, ten haplotypes were identified, eight of which have not previously been reported. Comparing these data with those from publicly available databases it appears that Ethiopian groups display much greater variation than do other populations (gene diversity using complete coding region haplotypes (non-synonymous variants only): Ethiopia = 0.17 ± 0.02; Rest of the World combined = 0.08 ± 0.03). Many haplotypes can be predicted to be of importance in the planning and provision of healthcare. In addition, Ethiopian populations exhibit most of the common variation observed elsewhere.

Chapter 3 focuses on extracting information about the recent evolutionary history of *CYP1A2* from the genetic variation observed in the Ethiopian ascertainment populations. A rooted *CYP1A2* haplotype network was produced which is consistent with humans evolving from non-human primates with the chimpanzee being their closest living relative. Of the populations used in the analysis, which included non-Africans and those with a recent African ancestry, the Ethiopian dataset contained haplotypes most similar to the chimpanzee. Haplotype networks revealed a varying level of conservation among the *CYP1A2* exons and showed that all haplotypes predicted to code for a protein with altered activity were observed in the external branches of the network. Intra-population gene diversities and inter-population genetic

distances were generally lowest for SNP categories expected to have greatest impact on protein structure. These results are consistent with the hypothesis that purifying selection has affected the allele frequencies of SNPs in *CYP1A2*, predicted to have greatest impact on protein structure, in humans. The time to most recent common ancestor of nine *CYP1A2* variants was estimated using variation in an AC microsatellite situated 5.6 kb downstream of the 3' end of *CYP1A2*. Coalescent date estimates place most variants into a period which pre-dates the expansion of modern humans out of Africa.

Chapter 4 addressed whether it would be possible to predict CYP1A2 functional variation in Ethiopian populations using variation known prior to this study. The test procedure was constructed utilising *CYP1A2* variant alleles, recorded by the CYP450 Allele Nomenclature Committee, which were assigned phenotypes from previously reported functional studies. The test's suitability for Ethiopia was investigated by applying it to the variation observed in the Ethiopian ascertainment populations. The diagnostic test was found to be inappropriate for Ethiopia since it did not account for the plethora of novel variation observed in the ascertainment samples. Ethiopia requires its own diagnostic test procedure initially built using the variation observed in the five Ethiopian ascertainment populations.

Chapter 5 focuses on variation within the transcription factor 7-like 2 gene (*TCF7L2*) which plays a role in the Wnt signalling pathway (Smith, 2007). *TCF7L2* alleles are associated with an increased risk of type 2 diabetes (T2D) in various populations (Weedon, 2007) but little is known about *TCF7L2* variability in Ethiopians. Varying levels of linkage disequilibrium were observed, in the Ethiopian ascertainment populations, between two SNPs (rs7903146 and rs12255372) which have been associated with an increased risk of T2D. Haplotypes defined by these two SNPs consequently captured almost all of the information on intra and inter-population diversity otherwise determined by analysing haplotypes defined by multiple SNPs. Both rs7903146 and rs12255372 were common in ~ 50 Ethiopian populations and haplotype distribution varied among groups suggesting that T2D susceptibility due to the effect of *TCF7L2* may also vary, with most groups likely having a West African *TCF7L2* risk for the disease and some having more of a European *TCF7L2* risk. The haplotypes could also effectively discriminate between Ethiopian populations in accordance with geography and linguistics states, and even had enough power to differentiate caste-like groups covered by the same self-identifying ethnic label.

Chapter 6 gives a general discussion of the outcomes of studies involving both *CYP1A2* and *TCF7L2*. Details of future work are also provided.

2 Variation in *CYP1A2* in Ethiopian populations

2.1 Introduction

2.1.1 Pharmacogenetics and cytochrome P450s

Pharmacogenetics is the study of individual genetic variation in response to drugs (Johnson, 2003; Weinshilboum, 2003; Wilke et al., 2007). When a drug is administered it must undergo the following processes (Weinshilboum, 2003):

1. Absorption
2. Distribution to site of action
3. Interaction with targets, for example enzymes and receptors
4. Metabolism
5. Excretion

Variation in each of these processes could potentially influence drug response (Weinshilboum, 2003). Variation in genes involved in drug transport, for example *ABCB1* (ATP binding cassette B1), and drug targets, such as the β_2- adrenoreceptor, are known (Evans and McLeod, 2003). However, variation in drug metabolism has received most attention. Over the past 20 years there has been increasing interest in drug metabolising enzymes, particularly cytochrome P450s (CYPs), as a literature search on the subject attests; over 46, 000 CYP articles have been published since 1989 (http://www.ncbi.nlm.nih.gov/pubmed/). Contributions to the field include studies examining functional genomics (Rezen et al., 2007), CYP genomic variation (Solus et al., 2004), the functional significance of mutations within CYP genes *in vivo* (Aklillu et al., 2003) and *in vitro* (Zhou et al., 2004), clinical implications of inter-individual CYP genomic variation (Kirchheiner and Seeringer, 2007), the evolutionary history of CYP genes (Heilmann et al., 1988; Goldstone et al., 2007), CYP roles in other species (Liang et al., 1996), links to cancer susceptibility (Agundez, 2004), Parkinson's disease (Tan et al., 2007), psychosomatic disease (Sugahara et al., 2009) and Alzheimer's (Van Ess et al., 2002).

CYP enzymes are highly potent *in vivo* oxidising agents, capable of catalysing the oxidative biotransformation of a diverse array of endogenous and exogenous substrates (Porter and Coon, 1991). They are a super-family of haem containing mono-oxygenases and are believed to have been synthesised in nature for over 3.5 billion years (Nebert and Russell, 2002). They were initially isolated from rat liver microsomes (Omura and Sato, 1964) but are widely distributed among other organisms, including bacteria, yeast, fungi and plants. Human CYPs are primarily membrane-associated proteins, located either in the inner membrane of mitochondria or in the endoplasmic reticulum of cells (Nelson, 2009). A large proportion of CYPs are found within most tissues of the body (e.g. intestine and lungs), but the majority are present in the liver (hepatic CYPs) where they are involved in the biosynthesis of cholesterol

and bile acids. A subset of CYPs also play important roles in steroidogenesis in the adrenals (adrenal CYPs) (Nelson, 2009). CYPs responsible for the metabolism of exogenous compounds are thought to have first evolved approximately 400-500 million years ago to allow animals to detoxify chemicals in plants, and excrete them from the body by making them more water soluble (Gonzalez and Gelboin, 1994). As a bi-product of this process, the enzymes also participate in the metabolic clearance of a vast array of clinical drugs.

The name **P450** refers to the 'pigment at 450 nm' formed by absorption of light at wavelengths near 450 nm when the haem iron is reduced (with sodium dithionite) and complexed to carbon monoxide. The enzymes incorporate one molecule of oxygen into the substrate and the other into water, and catalyse many types of reactions, including hydroxylation, epoxidation, N-dealkylation, O-dealkylation and S-oxidation, with hydroxylation being the most important. Many of the CYPs add a hydroxyl group to the substrate in Phase I of drug metabolism. In Phase II, this hydroxyl moiety then serves as the site for further modifications, which increase the solubility of the substrate enabling it to be excreted in urine (Nelson, 2009). Some examples of Phase II drug metabolising enzymes are glutathione S-transferases, N-acetyltransferases and sulfotransferases.

The Human Genome Project has revealed the presence of approximately 116 human CYP genes, of which 57 are active and 59 are pseudogenes (http://drnelson.utmem.edu/human.P450.table.html). The numerous CYP isoforms are classified as follows (Nelson et al., 1996):-

1. Families - >40% amino acid sequence homology is required for members to be in the same family. Families are numbered, e.g. *CYP1*, *CYP2*.

2. Subfamilies - >55% amino acid sequence homology is required for members to be in the same subfamily. Subfamilies are designated by letters, e.g. *CYP1A*, *CYP1B*, *CYP2C*.

3. Individual genes – genes are numbered. Similar function and high conservation is required for genes to have the same number, e.g. *CYP1A2*, *CYP1B1*, *CYP2C8*.

Human CYP enzymes in families 5-51 are important in the metabolism of endogenous substrates. Generally, they have high substrate specificity and have been relatively well conserved throughout evolution. Enzymes in CYP family 4, tend to metabolise fatty acids and related substrates, as well as interacting with some xenobiotics. Those in CYP families 1-3 have been conserved less well throughout evolution and exhibit important genetic polymorphisms. They generally have low substrate specificity, with many having overlapping substrate specificities (Ingelman-Sundberg, 2004). As a consequence of this, redundancy is evident within the CYP system (Gu et al., 1992). CYPs in families 1-3 are active in the

metabolism of a vast array of xenobiotic chemicals and are involved in as much as 75 % of phase I dependent drug metabolism. Approximately 40 % of this is initiated by genetically polymorphic CYP enzymes (Ingelman-Sundberg, 2004).

Mutations in a gene encoding a CYP enzyme can give rise to enzyme variants with higher, lower, or no activity (Masimirembwa and Hasler, 1997). Absence of enzyme activity may be caused by gene deletion, but can also occur when mutations lead to altered splicing, premature stop codons, damaged transcriptional initiation sites and amino acid changes which alter the structure/function of the protein (Pirmohamed and Park, 2003). CYP variants can lead to reduced enzyme activity (Lee et al., 2002) and altered activity; mutations in the active site (Oscarson et al., 1997) or changes in protein folding can alter the enzyme's substrate specificity (Johansson et al., 1994) for example. Increased enzyme activity can result from gene duplication, where individuals carry multiple copies of an active *CYP* gene (Johansson et al., 1993), but has also been associated with linked amino acid changes (Sakuyama et al., 2008). Alternatively, a variant in a *CYP* gene may have no direct effect on enzyme function but may be in linkage disequilibrium with a functional variant (Ingelman-Sundberg, 2001).

Variability in CYP activity can affect the efficacy and safety of drugs. Whilst individuals, known as extensive metabolisers (EMs), demonstrate normal levels of CYP activity, those with exceptionally high levels of CYP activity (ultra-rapid metabolisers, URMs) metabolise standard drug doses too quickly. As a result, the drug may be absent from blood serum or present at sub-therapeutic quantities with the consequence that these individuals require correspondingly higher doses of drugs to receive therapeutic benefit (Bertilsson et al., 1993). On the other hand, in individuals with exceptionally low levels of CYP activity (poor metabolisers, PMs), prodrugs may not be converted to the active form in sufficient quantity to be therapeutic or may accumulate in the body causing an adverse drug reaction (Coutts and Urichuk, 1999). For example, poor CYP2D6 metabolisers are unable to efficiently convert codeine (prodrug) to morphine (active form) and as a consequence may not experience pain relief (Desmeules et al., 1991). Ultrarapid CYP2D6 metabolisers may metabolize codeine too efficiently leading to morphine intoxication (Gasche et al., 2004).

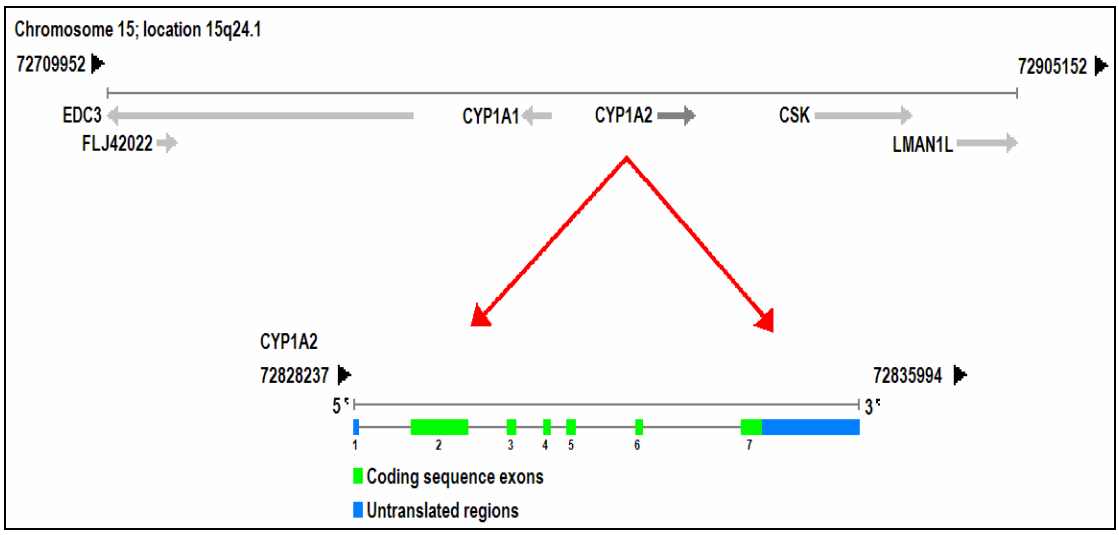
Inter-individual variability in drug clearance is thought to be common (Wilkinson, 2005) and CYP polymorphisms are frequently associated with adverse drug reactions (Ingelman-Sundberg, 2004). For instance, 59 % of drugs in adverse drug reaction studies are metabolised by polymorphic phase I dependent enzymes, 86 % of which are CYP enzymes (Phillips et al., 2001). Adverse drug reactions are a major burden on healthcare services and account for a great deal of morbidity and mortality, increased healthcare costs and pharmaceutical expenditure (Pirmohamed et al., 2004). For instance, adverse drug reactions caused 5 % of hospital admissions worldwide between 1966 and 1989 (Einarson, 1993) and 106, 000 deaths in the US in 1994 (Lazarou et al., 1998). Furthermore, it has been estimated that annual costs attributable to adverse drug reactions for a 700-bed hospital in the USA amount to $5.6 million (Bates et al., 1997).

2.1.2 Cytochrome P450 *1A2*

The human CYP1A subfamily, consisting of CYP1A1 and CYP1A2, is one of the clinically important members of the CYP enzyme superfamily, constituting approximately 15 % of the total CYP content in the liver (Shimada et al., 1994). While *CYP1A1* is expressed mainly in extrahepatic tissues (lung, skin, larynx and placenta), *CYP1A2* is expressed in the liver (Shimada et al., 1994). CYP1A2 detoxifies environmental xenobiotics in mammals and birds and is thought to have originated some 350 million years ago by duplication of *CYP1A1* (Heilmann et al., 1988; Liang et al., 1996). As a result, both CYP1A2 and CYP1A1 have 84 % amino acid sequence homology.

A cDNA corresponding to *CYP1A2* was isolated in 1987 (Jaiswal et al., 1987). Human *CYP1A2* is approximately 7.8 kb long with seven exons and six introns (Ikeya et al., 1989). Exon 1 and the downstream sequence of exon 7 are untranslated regions. The gene has only one transcript which is translated into a protein of 515 amino acid residues (NCBI36, http://www.ncbi.nlm.nih.gov/). The active site is thought to incorporate amino acids C458 and F451 in exon 7 and T321 in exon 4 (Sansen et al., 2007). *CYP1A2* is mapped to the positive strand of the long arm of chromosome 15 at 15q24.1 at chromosomal location 15:72,828,237-72,835,994 (NCBI36, http://www.ncbi.nlm.nih.gov/). *CYP1A2* lies approximately 25.5 kb upstream of the c-src tyrosine kinase (*CSK*) gene and is orientated head-to-head with *CYP1A1*, which is on the negative strand. *CYP1A2* and *CYP1A1* are separated by a 23.3 kb spacer region whose role in regulating one or other of the genes, or in governing the expression of both genes simultaneously, is not yet understood (Jiang et al., 2005). A schematic diagram of *CYP1A2* on chromosome 15 is shown in figure 2.1.

Figure 2.1 The location and structure of *CYP1A2*



CYP1A2 is responsible for the oxidative metabolism of a wide variety of drugs such as clozapine (Bertilsson et al., 1994; Fang et al., 1998), flutamide (Shet et al., 1997), lidocaine (Orlando et al., 2004) and caffeine (Butler et al., 1989). The enzyme is also involved in the

biotransformation of endogenous compounds such as bilirubin (Zaccaro et al., 2001) and melatonin (Facciola et al., 2001; Hartter et al., 2001), as well as the metabolic activation of procarcinogens such as aromatic, heterocyclic amines, polycyclic aromatic hyrocarbons and aflatoxin B1 (Butler et al., 1989; Boobis et al., 1994; Gallagher et al., 1994; Eaton et al., 1995). CYP1A2 is also induced by a number of products including cigarette smoke (Rasmussen et al., 2002) and 3-methylcholanthrene (Quattrochi et al., 1994). CYP1A2 is induced by 3-methylcholanthrene through binding to the aryl hydrocarbon receptor complex, which then binds to the xenobiotic responsive element located in the 5' region of the gene (Quattrochi et al., 1994). Several other transcription factor binding sites, associated with the regulation of *CYP1A2* expression, have also been identified in the 5' region, including AP-1, E-box proteins, and HNF-1 (Chung and Bresnick, 1995; Chung and Bresnick, 1997; Quattrochi et al., 1998; Pickwell et al., 2003). Furthermore, GC and CCAAT boxes have been identified approximately 940 nucleotides upstream of the translational start site (Chung and Bresnick, 1995). The enzyme also has many inhibitors including compounds found in oral contraceptives (Abernethy and Todd, 1985; Balogh et al., 1995). A summary of CYP1A2 substrates, inducers and inhibitors is shown in figure 2.2. Caffeine is frequently used as a substrate in CYP1A2 phenotype studies (Kalow and Tang, 1993; Fuhr et al., 1996) but theophylene (Sarkar et al., 1992) and melatonin (Hartter et al., 2001) are also used.

Figure 2.2 Known CYP1A2 substrates, inhibitors and inducers (derived from http://medicine.iupui.edu/flockhart/table.htm)

Substrates	**Inhibitors**	**Inducers**
Acetaminophen	amiodarone	beta-naphthoflavone
amitriptyline	cimetidine	broccoli
caffeine	ciprofloxacin	brussel sprouts
clomipramine	fluoroquinolones	char-grilled meat
clozapine	fluvoxamine	insulin
cyclobenzaprine	furafylline	methylcholanthrene
estradiol	interferon	modafinil
fluvoxamine	methoxsalen	nafcillin
haloperidol	mibefradil	omeprazole
imipramine		tobacco smoke
mexiletine		
naproxen		
olanzapine		
ondansetron		
phenacetin		
propranolol		
(R)warfarin		
riluzole		
ropivacaine		
tacrine		
theophylline		
tizanidine		
verapamil		
zileuton		
zolmitriptan		

CYP1A2 knockout mice develop normally but show deficient drug metabolism (Liang et al., 1996). Human *CYP1A2* mRNA expression varies up to 40 fold (Schweikl et al., 1993), and the *in vivo* activity of the enzyme may differ up to 60 fold (Eaton et al., 1995) among individuals. These differences may affect drug efficacy and safety as well as susceptibilities to cancer caused by procarcinogens in humans (Gunes and Dahl, 2008). Some of this variability may be due to environmental factors, such as tobacco smoke, exercise and/or diet. On the other hand,

it has been postulated that up to 75 % of the difference is due to genetic variation among individuals (Kendler and Prescott, 1999; Rasmussen et al., 2002). In addition, gender related differences in CYP1A2 activity have also been reported (Parkinson et al., 2004).

2.1.2.1 Known *CYP1A2* genomic variation prior to this study

To date, excluding data reported in this study, 125 allelic variants have been reported within *CYP1A2* (from exon 1 to exon 7) and a further 47 variants have been found within 3000 bases either side of the gene. The 5' region has 30 variants whilst the 3' region has 17 (see table 2.1 for a summary of the numbers and types of mutations). A comprehensive review of all these *CYP1A2* variants is set out in table 2.2 (please see CD); data regarding location, ancestral alleles, haplotypes, allele frequencies in different populations and functional studies are provided.

Table 2.1 Known *CYP1A2* variants prior to this study (data compiled from NCBI build 129 (http://www.ncbi.nlm.nih.gov/), *CYP1A2* allele nomenclature web page (www.cypalleles.ki.se/), NIEHS SNPs programme (www.egp.gs.washington.edu/data/CYP1A2/) and Pharmacogenetics Knowledge Base (www.pharmgkb.org/)).

Location	Substitutions	Transitions	Transversions	Insertion /Deletions	STR	Splice site	Synonymous	Non synonymous
5' end	28	17	11	2	0			
3' end	16	13	3	1	0			
3' UTR	13	9	4	4	0			
Intron 1	17	9	8	1	0	0		
Intron 2	7	6	1	0	0	0		
Intron 3	4	2	2	0	0	0		
Intron 4	6	5	1	0	0	0		
Intron 5	15	10	5	1	1	0		
Intron 6	12	8	4	2	0	1		
Exon 1 (5' UTR)	0	0	0	0	0			
Exon 2	22	13	9	0	0		6	16
Exon 3	7	5	2	0	0		3	4
Exon 4	1	1	0	0	0		0	1
Exon 5	3	2	1	0	0		0	3
Exon 6	2	2	0	0	0		1	1
Exon 7	7	6	1	0	0		1	6

The vast majority of variants are Single Nucleotide Polymorphisms (SNPs) which occur within the non-coding regions of the gene (table 2.1). Eleven variants are Deletion Insertion Polymorphisms (DIPs), none of which are found in the coding region, and one is a Short Tandem Repeat (STR) in intron 5. A total of 67 variants lie within the introns, one of which causes a splicing defect in intron 6. A total of 42 SNPs occur within the coding exons, over 70% of which are non-synonymous changes. To date, no variation has been reported in exon 1 (5' UTR), no more than seven variants have been found in any of exons 3, 4, 5 and 6, yet more than 20 variants have been found in each of exons 2 and 7 (including 3' UTR). Note that none of the variants are found in what is thought to be the active site of the protein. In addition, no copy number variation or gene conversion has been reported in *CYP1A2*. As far as linkage

disequilibrium (LD) across *CYP1A2* is concerned, patterns vary between HapMap populations (Yoruba, Europeans, Chinese and Japanese) and the gene is in one LD block in Europeans and Chinese but divided into different LD blocks in Yoruba and Japanese (http://www.hapmap.org/). LD was however high across *CYP1A2* in another Japanese cohort (Soyama et al., 2005). In a study involving Africans, Asians and Caucasians, *CYP1A1* and *CYP1A2* were each in their own block with high LD, and each block showed few historic recombination events. A recombination hotspot was identified in the middle of the spacer region which separates the two genes (Jiang et al., 2005).

Tables 2.3, 2.4 and 2.5 summarise the distribution (known prior to this study) of allele frequencies of *CYP1A2* variants, found above 1 % in at least one population, in the extragenic (5' and 3' regions), intronic and exonic regions respectively. It is clear that differences in allele frequencies, across the entire *CYP1A2* gene, exist among the study populations.

To date, prior to this study, 36 *CYP1A2* haplotypes (table 2.6), including 21 subtypes, have been named by the Human Cytochrome P450 Nomenclature Committee (www.cypalleles.ki.se/*CYP1A2*.htm). The rules for naming CYP haplotypes are as follows: The gene and haplotype are separated by an asterisk followed by Arabic numerals (e.g. *CYP1A2**1, *CYP1A2**2). A unique haplotype contains nucleotide changes which affect transcription, splicing, translation, posttranscriptional or posttranslational modifications or result in at least one amino acid change. Additional nucleotide changes are described by letters (upper-case Roman) such that where silent mutations occur, or where mutations are present in regulatory regions or introns with an unknown function, the haplotype name conforms to the closest functionally characterised allele by subgroup allocations e.g. *CYP1A2**1B, *1C and *1D (www.cypalleles.ki.se/criteria.htm). Most attention has been paid to intron 1 and the 5' region of the gene, due to this region's potential regulatory role in gene expression. A summary of the distribution of *CYP1A2* haplotypes, in the different study populations reported so far, is set out in table 2.6. Differences in haplotype frequencies among worldwide populations are evident.

2.1.2.2 CYP1A2 functional variation prior to this study

The associated functional status of each *CYP1A2* haplotype is also summarised in table 2.6. *CYP1A2**1J (Aklillu et al., 2003), *9, *10, *12, *13 and *14 (Murayama et al., 2004) were shown to have an enzyme activity similar to that of the 'wild type' *in vitro* (*1A haplotype). This was replicated *in vivo* for *CYP1A2**1J in the same study using caffeine as a phenotyping probe (substrate used to detect amount of CYP1A2 activity) (Aklillu et al., 2003). In contrast to this however, *CYP1A2**6 has been designated as being non-functional because the recombinant DNA expression (human *CYP1A2* expression from non-human cells) of the *6 variant, in

Table 2.3 The distribution of allele frequencies of *CYP1A2* variants, found above 1 % in at least one population, in the extragenic region (5' and 3' regions)

Population	Variant allele frequency																			Reference
	-3860A	-3598T	-3584G	-3113A	-2909T	-2847C	-2834T	-2808C	-2777A	-2733T	-2467-	-2103A	-1944A	-1804G	-1708C	-1051C	7352A	7573T	7864A	7894A	9540A	
African														0.110	0.160							Jiang et al., 2005
African American																	0.035		0.167	0.000		NIEHS SNPs[1]
				0.110		0.109																Perlegen-AFD[2]
	0.396				0.042	0.146	0.000	0.000	0.021	0.021	0.563	0.021	0.021	0.146	0.146			0.000	0.167			SNP500 Cancer[3]
	0.336					0.092																SNP500 Cancer HDP[3]
																0.000						Solus et al., 2004
Asian																	0.000		0.125	0.042		NIEHS SNPs[1]
														0.060	0.080							Jiang et al., 2005
Caucasian				0.017		0.000		0.000													0.000	CSHL HapMap[2]
				0.000		0.000																Perlegen-AFD[2]
	0.081				0.000	0.083	0.000	0.000	0.000	0.000	0.241	0.000	0.000	0.081	0.081			0.000	0.484			SNP500 Cancer[3]
	0.000					0.000																SNP500 Cancer HDP[3]
																0.030						Solus et al., 2004
														0.010	0.030							Jiang et al., 2005
Chinese								0.033														ABI-AoD[2]
				0.091		0.089		0.023													0.000	CSHL HapMap[2]
	0.220																					Han et al., 2002
				0.120		0.146																Perlegen-AFD[2]
																0.000						Solus et al., 2004
European																	0.000		0.595	0.000		NIEHS SNPs[1]
Hispanic																	0.000		0.310	0.000		NIEHS SNPs[1]
	0.196				0.000	0.068	0.000	0.000	0.000	0.000	0.250	0.022	0.000	0.045	0.045			0.025	0.435			SNP500 Cancer[3]
Japanese								0.011														ABI-AoD[2]
											0.420											Chida et al., 1999
				0.033		0.044		0.011													0.000	CSHL HapMap[2]
	0.230																					Nakajima et al., 1999
																0.000						Solus et al., 2004
	0.236	0.032	0.170	0.026		0.032		0.016			0.438											Soyama et al., 2005
Korean	0.270			0.030							0.707											Ghotbi et al., 2007
Native American/Hispanic	0.471					0.031																SNP500 Cancer HDP[3]
Pacific Rim	0.239				0.000	0.042	0.021	0.000	0.000	0.000	0.625	0.000	0.000	0.042	0.042			0.000	0.125			SNP500 Cancer[3]
	0.145					0.074																SNP500 Cancer HDP[3]
South East Asian																0.000						Solus et al., 2004
Swedish	0.010			0.020							0.193											Ghotbi et al., 2007
Yoruba				0.100		0.125		0.000													0.067	CSHL HapMap[2]
																	0.000		0.000	0.000		NIEHS SNPs[1]

[1] http://egp.gs.washington.edu/data/cyp1a2/
[2] NCBI build 129 http://www.ncbi.nlm.nih.gov/
[3] http://snp500cancer.nci.nih.gov/

Table 2.4 The distribution of allele frequencies of *CYP1A2* variants, found above 1 % in at least one population, in the intronic region

Population	Variant allele frequency																																Reference
	-739G	-733C	-729T	-662G	-569A	-505A	-367T	-349G	-178A	-163A	-61G	850C	869C	894T	1202T	1280G	1352A	1590T	1649T	1669T	2159A	2169A	2232C	2321C	2570A	2597G	2645G	2646T	2694C	3069C	3533A	3605G	3613C	4310A	4598T	5010T	
African	0.160								0.060	0.430					0.000						0.010												0.130				Jiang et al., 2005
African American										0.644																											ABI-AoD[2]
					0.033	0.033			0.033	0.654	0.000		0.000		0.000				0.000	0.000	0.133												0.107	0.033	0.167		NIEHS SNPs[1]
	0.109									0.543					0.000						0.087																Perlegen-AFD[2]
	0.146	0.000	0.000	0.000				0.021	0.000	0.771		0.021		0.000	0.000	0.021	0.000				0.167	0.000	0.000	0.000	0.042	0.042	0.000	0.000	0.000	0.021	0.000	0.021	0.130			0.000	SNP500 Cancer[3]
	0.095		0.000							0.561					0.000						0.086			0.020							0.000		0.079				SNP500 Cancer HDP[3]
	0.060	0.000	0.000															0.000			0.380			0.000	0.130			0.000	0.000				0.060			0.000	Solus et al., 2004
	0.133		0.011																											0.000						PharmGKB AB-DME[4]
Asian					0.000	0.000			0.000	0.625	0.000		0.000		0.021				0.021	0.000	0.130												0.021	0.000	0.229		NIEHS SNPs[1]
	0.067		0.000																											0.000						SNP500 Cancer HapMap[3]
	0.080								0.000	0.440					0.080						0.180												0.080				Jiang et al., 2005
British										0.660																											Sachse et al., 2003
Caucasian										0.783																											ABI-AoD[2]
	0.000		0.000							0.692					0.000						0.644			0.033								0.000				CSHL HapMap[2]
	0.000									0.708					0.000						0.688																Perlegen-AFD[2]
	0.081	0.000	0.032	0.000				0.000	0.000	0.726		0.000		0.016	0.081	0.000	0.032				0.500	0.016	0.016	0.065	0.000	0.000	0.016	0.000	0.000	0.000	0.000	0.000	0.081			0.032	SNP500 Cancer[3]
	0.000		0.000							0.659					0.000						0.660			0.030							0.000		0.000				SNP500 Cancer HDP[3]
	0.060	0.000	0.010															0.010			0.390			0.030	0.000			0.010	0.000				0.060			0.030	Solus et al., 2004
	0.005																																				Chevalier et al., 2001
	0.000		0.000																											0.000						SNP500 Cancer HapMap[3]
	0.056		0.000																											0.000						PharmGKB AB-DME[4]
																														0.005						Allorge et al., 2003
	0.030								0.000	0.300					0.020						0.640												0.030				Jiang et al., 2005
Chinese										0.689											0.167																ABI-AoD[2]
	0.083		0.023							0.667					0.091						0.122			0.151								0.061				CSHL HapMap[2]
										0.660																											Han et al., 2002
	0.174									0.688					0.125						0.167																Perlegen-AFD[2]
	0.100	0.100	0.000															0.000			0.150			0.150	0.000			0.000	0.000				0.100			0.000	Solus et al., 2004
	0.140		0.000																											0.000						PharmGKB AB-DME[4]
Egyptian	0.030									0.680																											Hamdy et al., 2003
Ethiopian	0.075		0.030							0.496																											Aklillu et al., 2003
European					0.000	0.000			0.000	0.659	0.023		0.000		0.000				0.000	0.000	0.625												0.000	0.000	0.024		NIEHS SNPs[1]
										0.712																											SEQUENOM-CEPH[2]
German										0.680																											Sachse et al., 1999
Hispanic					0.000	0.000			0.000	0.857	0.000		0.023		0.023				0.000	0.000	0.325												0.023	0.023	0.095		NIEHS SNPs[1]
	0.043	0.000	0.000	0.043				0.023	0.000	0.696		0.000		0.000	0.000	0.000	0.000				0.435	0.000	0.000	0.022	0.000	0.000	0.000	0.022	0.000	0.000	0.000	0.000	0.022			0.065	SNP500 Cancer[3]
Italian										0.670																											Pavanello et al., 2005
Japanese										0.644											0.244																ABI-AoD[2]
	0.082									0.613																											Chida et al., 1999
	0.023		0.000							0.614					0.033						0.182			0.122								0.012				CSHL HapMap[2]
	0.000	0.050	0.000															0.000			0.250			0.200	0.000			0.000	0.000			0.000			0.000	Solus et al., 2004
	0.032	0.020	0.032				0.012			0.628									0.002		0.194			0.170						0.000		0.026				Soyama et al., 2005
	0.044		0.000																											0.011						PharmGKB AB-DME[4]
Korean	0.027		0.000							0.627																											Ghotbi et al., 2007
Native American/Hispanic	0.032		0.000							0.660					0.031						0.071			0.041							0.000		0.031				SNP500 Cancer HDP[3]
Pacific Rim	0.042	0.083	0.000	0.000				0.000	0.021	0.750		0.000		0.000	0.042	0.000	0.000				0.125	0.000	0.000	0.354	0.000	0.000	0.000	0.000	0.021	0.000	0.000	0.000	0.042			0.000	SNP500 Cancer[3]
	0.057		0.000							0.580					0.069						0.189			0.157							0.000		0.073				SNP500 Cancer HDP[3]
Saudi Arabian	0.059		0.036																																		Aklillu et al., 2003
South East Asian	0.000	0.000	0.000															0.000			0.150			0.200	0.000			0.000	0.050				0.000			0.000	Solus et al., 2004
Spanish	0.013		0.005																																		Aklillu et al., 2003
Swedish	0.023		0.003							0.714																											Ghotbi et al., 2007
Tanzanian										0.490																											Dandara et al., 2004
Yoruba	0.125		0.000							0.567					0.000						0.008			0.025								0.059				CSHL HapMap[2]
					0.000	0.042			0.000	0.583	0.000		0.000		0.000				0.000	0.042	0.000												0.083	0.000	0.125		NIEHS SNPs[1]
	0.125		0.000																											0.000						SNP500 Cancer HapMap[3]
Zimbabwean										0.570																											Dandara et al., 2004

[1] http://egp.gs.washington.edu/data/cyp1a2/
[2] NCBI build 129 http://www.ncbi.nlm.nih.gov/
[3] http://snp500cancer.nci.nih.gov/
[4] http://www.pharmgkb.org/

Table 2.5 The distribution of allele frequencies of *CYP1A2* variants, found above 1 % in at least one population, in the exonic region

Population	Variant allele frequency																							Reference
	53G	217A	222T	310A	331T	393A	413A	613G	1460T	1513A	1514A	1559G	2116A	5090T	5112T	5168T	5342G	5347T	5521G	5654T	5890G	6021T	6324-	6537A	6674G	6685G	
African										0.000								0.010							0.190		Jiang et al., 2005
African American																		0.155									ABI-AoD[2]
		0.000		0.000	0.033			0.000	0.000	0.033					0.000			0.133	0.167	0.000		0.000	0.167	0.000	0.192		NIEHS SNPs[1]
																		0.091									Perlegen-AFD[2]
	0.000		0.000	0.000		0.000	0.000			0.146	0.000		0.000	0.000		0.000	0.000	0.167						0.000	0.146	0.021	SNP500 Cancer[3]
	0.013						0.000			0.118			0.000	0.000				0.047									SNP500 Cancer HDP[3]
	0.000									0.130		0.000		0.000				0.370	0.070								Solus et al., 2004
	0.000						0.000			0.067			0.000	0.000	0.000												PharmGKB AB-DME[4]
Asian		0.000		0.000	0.000			0.000	0.000	0.000					0.028			0.125	0.229	0.000		0.000	0.000	0.000	0.222		NIEHS SNPs[1]
	0.000						0.000			0.000			0.000	0.000													SNP500 Cancer HapMap[3]
										0.080								0.150							0.250		Jiang et al., 2005
British																		0.390									Sachse et al., 2003
Caucasian																		0.652									ABI-AoD[2]
																		0.642									CSHL HapMap[2]
																		0.630									Perlegen-AFD[2]
	0.000		0.032	0.016		0.000	0.000			0.000	0.000		0.000	0.000		0.000	0.000	0.500						0.000	0.177	0.000	SNP500 Cancer[3]
	0.000						0.000			0.000			0.000	0.000				0.617									SNP500 Cancer HDP[3]
	0.000									0.000		0.030		0.020				0.370	0.080								Solus et al., 2004
													0.010	0.005				0.330									Chevalier et al., 2001
	0.000						0.000			0.000			0.000	0.000													SNP500 Cancer HapMap[3]
	0.000						0.022			0.000			0.000	0.000	0.000												PharmGKB AB-DME[4]
										0.020								0.600							0.090		Jiang et al., 2005
Chinese																		0.189									ABI-AoD[2]
																		0.122									CSHL HapMap[2]
																		0.174									Perlegen-AFD[2]
	0.000									0.000		0.000		0.000				0.150	0.220								Solus et al., 2004
	0.000						0.000			0.000			0.000	0.000	0.000												PharmGKB AB-DME[4]
European		0.000		0.023	0.000			0.000	0.023	0.000					0.000			0.625	0.025	0.000		0.000	0.000	0.000	0.026		NIEHS SNPs[1]
																		0.620			0.093						SEQUENOM-CEPH[2]
Hispanic		0.000		0.000	0.000			0.000	0.000	0.000					0.000			0.333	0.079	0.053		0.023	0.000	0.000	0.083		NIEHS SNPs[1]
	0.000		0.000	0.000		0.000	0.000			0.022	0.000		0.000	0.000		0.000	0.023	0.435						0.000	0.065	0.022	SNP500 Cancer[3]
Japanese																		0.244									ABI-AoD[2]
																		0.182									CSHL HapMap[2]
	0.000									0.000		0.000		0.000				0.250	0.200								Solus et al., 2004
											0.006		0.000	0.000	0.006			0.192	0.192								Soyama et al., 2005
	0.000						0.000			0.000			0.000	0.000	0.011												PharmGKB AB-DME[4]
											0.006				0.004												Murayama et al., 2004
Native American/Hispanic	0.000						0.000			0.000			0.000	0.000				0.043									SNP500 Cancer HDP[3]
Pacific Rim	0.000		0.000	0.000		0.021	0.000			0.000	0.042		0.000	0.000		0.021	0.000	0.125						0.000	0.413	0.000	SNP500 Cancer[3]
	0.000						0.000			0.000			0.000	0.000				0.192									SNP500 Cancer HDP[3]
South East Asian	0.000									0.000		0.000		0.000				0.150	0.200								Solus et al., 2004
Yoruba																		0.000									CSHL HapMap[2]
		0.042		0.000	0.000			0.042	0.000	0.125					0.000			0.000	0.091	0.000		0.000	0.125	0.045	0.150		NIEHS SNPs[1]
	0.008						0.000			0.110			0.000	0.000													SNP500 Cancer HapMap[3]

[1] http://egp.gs.washington.edu/data/cyp1a2/
[2] NCBI build 129 http://www.ncbi.nlm.nih.gov/
[3] http://snp500cancer.nci.nih.gov/
[4] http://www.pharmgkb.org/

Table 2.6 The distribution of *CYP1A2* haplotypes in different study populations and their associated function

CYP1A2 haplotype	Nucleotide changes	Amino acid change	Enzyme activity	Reference for functional study	Haplotype frequencies						
					Ethiopian (n=346)	Saudi Arabian (n=272) (Aklillu et al., 2003)	Spanish (n=234)	Swedish (n=386) (Ghotbi et al., 2007)	Korean (n=100)	Japanese (n=500) (Soyama et al., 2005)	French Caucasian (n=200) (Chevalier et al., 2005)
*1A	None (unmutated for all alleles)		Wild type function	Designated wild type function by Allele Nomenclature Committee	0.399 (only intron 1 analysed - could be *1B for example)			0.244 (only intron 1 and enhancer region analysed - could be *1B for example)	0.217 (only intron 1 and enhancer region analysed - could be *1B for example)		0.635
*1B	5347T>C	N516N	Undetermined							0.348 (incorrectly reported as *1A)	0.330
*1C	-3860G>A		Reduced function	Nakajima et al., 1999							
*1D	-2467delT		Undetermined					0.034	0.153		
*1E	-739T>G		Undetermined								
*1F	-163C>A		Undetermined (controversial)	Sachse et al., 1999; Han et al., 2001; Han et al., 2002; Nordmark et al., 2002; Shimoda et al., 2002; Aklillu et al., 2003; Castorena-Torres et al., 2005	0.496			0.567	0.077	0.004	
*1G	-739T>G; 5347T>C	N516N	Undetermined								0.005
*1H	1570A>C; 5347T>C	A317A; N516N	Undetermined								0.005
*1J	-739T>G ; -163C>A		Wild type function	Aklillu et al., 2003	0.075	0.059	0.013				
*1K	-739T>G; -729C>T; -163C>A		Reduced function	Aklillu et al., 2003	0.030	0.036	0.005	0.003	0.000		
*1L	-3860G>A; -2467delT;-163C>A; 5347T>C	N516N	Undetermined					0.008	0.267	0.230	
*1M	-163C>A; 2159G>A		Undetermined							0.174	
*1N	-3594T>G; -2467delT;-163C>A; 2321G>C; 5521A>G; 5347T>C	N516N	Undetermined							0.108	
*1P	-3594T>G; -2467delT;-733G>C; 163C>A; 2321G>C; 5521A>G; 5347T>C	N516N	Undetermined							0.020	
*1Q	-2808A>C; -163C>A; 2159G>A		Undetermined							0.016	
*1R	-3594T>G; -2467delT;-367C>T; -163C>A; 2321G>C; 5521A>G; 5347T>C	N516N	Undetermined							0.010	
*1S	-3053A>G; 5347T>C	N516N	Undetermined							0.002	
*1T	-2667T>G; 5347T>C	N516N	Undetermined							0.002	
*1U	678C>T; 5347T>C	F226F; N516N	Undetermined							0.002	
*1V	-2467delT; -163C>A		Undetermined	Ghotbi et al., 2007				0.123	0.260		
*1W	-3113G>A; -2467delT; -739T>G; -163C>A		Undetermined (controversial)	Chen et al., 2005; Jiang et al., 2006; Ghotbi et al., 2007				0.021	0.027		
*2	63C>G	F21L	Undetermined							0.000	
*3	2116G>A; 5347T>C	D348N; N516N	Undetermined	Zhou et al., 2004						0.000	0.010
*4	2499A>T	I386F	Undetermined	Zhou et al., 2004						0.000	0.005
*5	3496G>A	C406Y	Undetermined	Zhou et al., 2004						0.000	0.005
*6	5090C>T	R431W	Non-functional	Zhou et al., 2004						0.000	0.005
*7	3533G>A	Splicing defect	Reduced function	Allorge et al., 2003						0.000	
*8	5166G>A; 5347T>C	R456H; N516N	Reduced function	Saito et al., 2005						0.004	
*9	248C>T	T83M	Wild type function	Murayama et al., 2004						0.004	
*10	502G>C	E168Q	Wild type function	Murayama et al., 2004						0.002	
*11	558C>A	F186L	Reduced function	Murayama et al., 2004						0.002	
*12	634A>T	S212C	Wild type function	Murayama et al., 2004						0.004	
*13	1514G>A	G299S	Wild type function	Murayama et al., 2004						0.006	
*14	5112C>T	T438I	Wild type function	Murayama et al., 2004						0.006	
*15	125C>G; 5347T>C	P42R; N516N	Reduced function	Saito et al., 2005						0.002	
*16	2473G>A; 5347T>C	R377Q; N516N	Reduced function	Saito et al., 2005						0.002	

Escherichia Coli, led to no detectable levels of CYP1A2 (Zhou et al., 2004). *CYP1A2**1C, *1K, *7, *8, *11, *15 and *16 all produced proteins with reduced CYP1A2 activity compared to the 'wild type'. The -3860 G>A mutation in *1C caused a decrease in the inducibility of the enzyme in Japanese smokers (Nakajima et al., 1999). *CYP1A2**1K reporter constructs showed significantly reduced inducibility *in vitro* and subjects with *CYP1A2**1K showed significantly decreased CYP1A2 activity using caffeine as the phenotyping probe (Aklillu et al., 2003). The -729 C>T mutation in intron 1 was thought to be the causal variant of this reduced phenotype. The point mutation 3533 G>A in *CYP1A2**7 caused a splicing defect in intron 6 and is likely to be the reason why carriers of this variant had reduced CYP1A2 activity when phenotyped with caffeine (Allorge et al., 2003). The F186L amino acid substitution in *CYP1A2**11 had no effect on protein expression but showed a markedly decreased catalytic activity *in vitro* in V79 transfected hamster cells; 12 % of the wild type's capacity for phenacetin *O*-deethylation and 28 % for 7-ethoxyresorufin *O*-deethylation (Murayama et al., 2004). *CYP1A2**8, *15 and *16 showed reduced protein expression levels and less than 1 % of the 7-ethoxyresorufin *O*-deethylation and 4 % of the phenacetin *O*-deethylation capacity compared with the *1A variant in V79 hamster cells (Saito et al., 2005). None of these variants have however been functionally assessed in humans. The recombinant DNA expression of *CYP1A2* variants *3, *4 and *5 led to varying levels of decreased expression and catalytic activities, and altered substrate specificity for phenacetin and heterocyclic amines (Zhou et al., 2004). Their function is still however classified as being "undetermined" (personal communication, P. David Josephy[2]). *CYP1A2**2 has only been found in a Chinese cohort at a frequency below 1 % (n = 1) (Huang et al., 1999). The functional significance of this variant has not been evaluated. The -163 C>A mutation in intron 1 (*CYP1A2**1F) was associated with higher enzyme inducibility by smoking (Sachse et al., 1999), although this association was not always replicated (Nordmark et al., 2002; Aklillu et al., 2003). Likewise, the -3113 G>A variant in *CYP1A2**1W was shown to alter CYP1A2 activity in non-smoking Chinese individuals. However, this was not replicated in Caucasians, Asians or Africans (Jiang et al., 2006; Ghotbi et al., 2007). *CYP1A2**1B, *1D, *1E, *1G, *1H, *1L, *1M, *1N, *1P, *1Q, *1U, *1R, *1S, *1T, *1U, *1V and *1W are haplotypes whose functional significance remains 'undetermined'.

2.1.3 Rationale of study

Little investigation of *CYP1A2* has been undertaken in the Ethiopian population to date. Researchers at the Karolinska Institute in Stockholm (Aklillu et al., 2003) have carried out *CYP1A2* genotype and phenotype studies in 100 Ethiopians from Ethiopia and 73 living in Sweden. However this study only sequenced the gene in 12 individuals, genotyping was restricted to intron 1 and the sample set was of a mixed Ethiopian origin from the Oromo, Amhara, Tigriyan and Gurage ethnic groups. A group at the University of Cincinnati Medical Centre have also carried out *CYP1A2* genotype studies in Ethiopians as part of a wider study (Jiang et al., 2006). However, they only genotyped six SNPs in six Ethiopians, whose ethnicity

[2] P.David Josephy is a co-author of Zhou et al., 2004

is not recorded. Despite this lack of information however, many pharmaceuticals, which may be metabolised by CYP1A2, are administered in Ethiopia. For example, both primaquine and praziquantel are used as the first line of treatment for malaria and schistosomiasis respectively (Federal Democratic Republic of Ethiopia Ministry of Health, 2004) and CYP1A2 is thought to be involved in their metabolism (Li et al., 2003). Furthermore, coffee was first domesticated for human use in Ethiopia (Anthony et al., 2002) and is an integral part of modern Ethiopian culture. The intake of caffeine (a well known CYP1A2 substrate) is consequently widespread in Ethiopia.

2.1.4 Aims

1. Characterise the distribution of genetic variation in exons and their flanking intronic regions in *CYP1A2* in the Ethiopian ascertainment populations.

2. Determine the extent of variation in comparison to what is already known.

3. Identify haplotypes common in Ethiopia and the five ethnic groups studied.

4. Analyse Ethiopian data in the context of *CYP1A2* variation reported by the Institute of Environmental Health Sciences (NIEHS) SNPs programme in African Americans, Yoruba, Europeans, Hispanics and East Asians.

2.2 Methods

2.2.1 Ethiopian samples

DNA samples were collected from males, 18 years old or older, unrelated at the paternal grandfather level. All samples were collected anonymously with informed consent. Sociological data, including age, current residence, birthplace, self-declared ethnic identity and religion of the individual and of the individual's father, mother, paternal grandfather and maternal grandmother were also collected.

Samples from the following ethnic groups were chosen for an 'Ethiopian ascertainment panel': Afar (n = 76), Amhara (n = 77), Anuak (n = 76), Maale (n = 76) and Oromo (n = 76). Afar were collected from Dubti (11.74 ºN, 41.09 ºE) and Asayta (11.56 ºN, 41.44 ºE) in Afar, Amhara and Oromo from Addis Ababa (9.03 ºN, 38.70 ºE) and Jimma (7.67 ºN, 36.83 ºE), Anuak from the Gambela region (including Gog (7.58 ºN, 34.50 ºE), Itang (8.20 ºN, 34.27 ºE) and Akobo (7.82 ºN, 33.03 ºE)) and Maale from Jinka (5.65 ºN, 36.65 ºE) in the Bako Gazer woreda in South Omo.

2.2.2 NIEHS samples

The NIEHS in the USA have set up the Environmental Genome Project in an effort to examine the relationships between environmental exposures, inter-individual sequence variation in human genes and disease risk in U.S. populations. The NIEHS SNPs Programme at the University of Washington (http://egp.gs.washington.edu/) systematically identifies and genotypes SNPs in environmental response genes as part of the wider study. Human *CYP1A2* has been sequenced by the NIEHS SNPs Programme in a total of 95 HapMap samples from the following ethnicities: 12 Yoruba, 15 African American, 22 European, 22 Hispanic and 24 East Asian (12 Japanese and 12 Han Chinese). The genotype data is publicly available and was incorporated in the analyses of this study to place the Ethiopian data in a worldwide context.

2.2.3 DNA extraction from buccal swabs

Buccal swabs were stored in 1 ml of preservative solution (0.05 M Ethylenediaminetetraacetic acid (EDTA), PH 8.0, 0.5 % Sodium Dodecyl Sulfate (SDS). DNA was extracted using a Phenol/Chloroform DNA extraction protocol.

To begin with, 40 µl of 10 mg/ml proteinase K was added to 20 ml of sterile distilled water. 0.8 ml of this solution was added to the 1.5 ml tube containing the buccal swab immersed in EDTA/SDS solution. Following incubation at 56 ºC for 1-3 hours, 0.8 ml of the mixture was added to a microfuge tube containing 0.6 ml of phenol/chloroform (1:1) mix. The sample was mixed and centrifuged for 10 minutes at 2240 xg. The resultant aqueous (upper) phase (layer) was transferred to a new microfuge tube containing 0.6 ml of chloroform and 30 µl of 5 M NaCl. The sample was mixed and centrifuged for 10 minutes at 2240 xg. The resultant aqueous (upper) phase (layer) was transferred to a new microfuge tube containing 0.7 ml of chloroform, the sample was mixed and centrifuged for 10 minutes at 2240 xg. The resultant aqueous (upper) phase (layer) was transferred to a screw-top microfuge tube (used for long term storage of DNA) containing 0.7 ml of isopropanol. The sample was mixed and centrifuged for 13 minutes at 2240 xg. The resultant supernatant was carefully (to avoid dislodging the DNA from the walls of the tube) discarded and the tube was inverted at 45 ˚ for one minute in order to drain off any remaining supernatant. 0.8ml of 70 % Ethanol was then added to the screw-top microfuge tube and the mixture was centrifuged for 10 minutes at 2240 xg. The resultant supernatant was carefully discarded and the tube was inverted at 45 ˚ for 20 minutes in order to drain off any remaining supernatant. 200 µl of TE (pH 9.0) was then added to the microfuge tube. The mixture was then incubated at 56 ºC for 10 minutes mixing occasionally. The resulting DNA with TE mixture was then stored upright at -20 ºC.

2.2.4 Amplification of *CYP1A2* *

All *CYP1A2* exons and flanking introns were amplified in six amplicons and in two PCR rounds. Primer details for first and second round PCRs are shown in table 2.7. Each DNA amplicon was amplified in a separate first round PCR and in 10 µl reaction volumes containing 1 ng of template DNA, 1 µM of each primer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.2 µl of Platinum Taq (Invitrogen) and $MgCl_2$ - free buffer supplied with the Taq. Cycling parameters were: 5 minutes of pre-incubation at 96 ºC, followed by 25 cycles of 30 seconds at 96 ºC, 30 seconds at 45 ºC and 45 seconds at 45 ºC, with a final elongation step for 7 minutes at 72 ºC. 1 µl of each of the purified first round PCR products was amplified in a second round PCR. Reaction volumes and compositions were the same as those used in the first round PCR, with the exception of the DNA template and primers. Cycling parameters were: 5 minutes of pre-incubation at 96 ºC, followed by 25 cycles of 30 seconds at 96 ºC, 30 seconds at 50 ºC and 45 seconds at 72 ºC, with a final elongation step for 7 minutes at 72 ºC. PCR products were purified using ExoSAP-IT (USB Corporation). Exonuclease I (Exo) removes residual single-stranded primers and any extraneous single-stranded DNA produced in the PCR. Shrimp Alkaline Phosphatase (SAP) removes remaining dNTPs from the PCR mixture. Briefly, 0.5 µl of enzyme mix was added to 1 µl of PCR product and incubated for 15 minutes at 37 ºC, followed by 15 minutes at 80 ºC.

2.2.5 Sequencing of *CYP1A2* *

Bidirectional sequencing was performed for each amplicon from the second round PCR with an overlap in the centre. Sequencing primers were the same as the primers used in the second round PCRs (table 2.7). Primers were designed to enable as much of the amplicon as possible to be sequenced in both directions. DNA was sequenced in 5 µl reaction volumes containing 1.5 µl of purified PCR template, 1 µl of BigDye termination mix v3.1 (Applied Biosystems (ABI), Warrington UK) and 1.5 µl of buffer supplied with the mix. Cycling parameters were: 2 minutes of pre-incubation at 96 ºC, followed by 25 cycles of 15 seconds at 96 ºC, 10 seconds at 50 ºC and 3 minutes at 60 ºC. Following standard ethanol precipitation, samples were then run on an ABI 3730 genetic analyser and analysed using Sequencher 4.7 software (Gene Codes Corporation, USA). All six amplicons were sequenced again in 20 samples in order to confirm sequencing. Methods highlighted with * were performed by Macrogen USA (Rockville MD, USA) (all other methods were performed by me). Macrogen USA sequencing was further confirmed by sequencing the coding sequence (cds) of exon 7 according to a different method (performed by me) described below.

2.2.6 Amplification of *CYP1A2* exon 7 (coding sequence)

A 531 bp region, containing the cds region of *CYP1A2* exon 7, was amplified using primers *CYP1A2*F CCTTCATTGCTTTCAAAGTGCC and *CYP1A2*R CTGCACTTGGCTAAAGCTGCT.

Table 2.7 First (yellow) and second (green) round PCR primers for the amplification of *CYP1A2*. Second round PCR primers were also used for sequencing.

Amplicon	First round PCR				Second round PCR and sequencing		
	Length (bp)	Amplicon location	Forward primer	Reverse primer	Length (bp)	Forward primer	Reverse primer
1	730	Exon 1 and flanking regions	GCTCCCTACCCTGAACCCTA	TCCATATACCCAAGGGACCA	646	TGGCCTATCCCCAAAGAGTCAC	AGTTCCCCCTACCCAGTGAC
2	1465	Exon 2 and flanking regions	TACCCAGCATGCATGCTG	CATTGCAGGACTCTGCTAGG	1315	CTACTCCAGCCCCAGAAGTG	GGCTTCAAGGATGAGGAAAC
3	1648	Exons 3, 4 and 5, introns 3 and 4 and part of introns 2 and 5	CTCTGGTGTCACGTTGCTT	TTAGCAAGATTGGAGGCCAA	1516	GGTGTTATTGGGAGGAAGGG	ACTGGGAGGGAGGGAATATG
4	1000	Exon 6 and flanking regions	CCCAAACGTTGTTCTAGTTATT	ATCACCTGTAACAAACGTCT	534	TGAAATTGCCTGCTTCTTGG	TGGCCATCCTAGTTGATTCC
5	1650	Part of intron 6 and the beginning of exon 7	TTAGCCGGATATGGTGCCTG	CGGTGGTTCATACCTGTTAAT	1302	TTAGCCGGATATGGTGCCTG	AATGTAAGTTAGGCTGGATGTG
6	1293	The end of exon 7 and part of the 3' near gene region	GAGTCACTACGCCTGGCTGA	TGCCTTTTGAGAATGGGACA	1243	GAGTCACTACGCCTGGCTGA	GACTGGATCCCTTTCTCCTTG

DNA was amplified in 96 well plates in 10 µl reaction volumes containing 1 ng of template DNA, 0.3 µM of each primer, 2 % dimethyl sulphoxide (DMSO), 0.13 units *Taq* DNA polymerase (HT Biotech, Cambridge, UK), 9.3 nM TaqStartTM monoclonal antibody (BD Biosciences Clontech, Oxford, UK), 200 µM dNTPs and reaction buffer supplied with the *Taq* polymerase. The cycling parameters were: 5 minutes of pre-incubation at 94 ºC, followed by 39 cycles of 1 minute at 94 ºC, 1 minute at 64 ºC and 1 minute at 72 ºC, with a final elongation step for 7 minutes at 72 ºC. The resultant PCR product was purified by mixing 30 µl of 1/3 water 2/3 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 PCR product and centrifuging the mixture at 2240 xg for 45 minutes. The resultant supernatant was discarded by inverting the PCR plate and centrifuging for 1 minute at 13 xg. 150 µl of 70 % Ethanol was added to each well and the plate was centrifuged at 2240 xg for 25 minutes. The resultant supernatant was discarded by inverting the PCR plate and centrifuging for 1 minute at 13 xg. Samples were dried for 5 minutes at 65 ºC and 30 µl of water was added to each purified PCR product. The solution was mixed gently and heated at 65 ºC for 5 minutes to resuspend the pellet.

2.2.7 Sequencing of *CYP1A2* exon 7 (coding sequence)

Bidirectional sequencing was performed using the same forward and reverse primers used in the first round PCR (*CYP1A2*F and *CYP1A2*R). DNA was sequenced in 96 well plates in 15 µl reaction volumes containing 6 µl of the purified PCR product, 2.4 pm of either primer, 0.75 µl of BigDye termination mix v1.1 (Applied Biosystems (ABI), Warrington UK) and 5 µl of Better Buffer (Applied Biosystems (ABI), Warrington UK). The cycling parameters were: 25 cycles of 96 ºC for 10 seconds, 55 ºC for 5 seconds and 60 ºC for 4 minutes. Sequencing reaction products were purified by mixing 80 µl of isopropanol with each product and leaving for 10 minutes at room temperature. The mixture was centrifuged for 45 minutes at 2240 xg and the resultant supernatant was discarded by inverting the PCR plate and centrifuging for 1 minute at 13 xg. 150 µl of 70 % isopropanol was added to each well and centrifuged for 15 minutes at 2240 xg. The resultant supernatant was discarded by inverting the PCR plate and centrifuging for 1 minute at 13 xg. Samples were then dried for 5 minutes at 65 ºC. Each sample was mixed with 10 µl of high purity (HiDi) formamide, heated for 4 minutes at 96 ºC and immediately cooled in ice. Samples were then run on an ABI 3100 genetic analyser and analysed using Sequencher 4.7 software (Gene Codes Corporation, USA). Macrogen USA sequencing (highlighted with *) was further confirmed by genotyping alleles at positions -163 (intron 1) and 2159 (intron 4), via TaqMan technology as described below.

2.2.8 Genotyping of -163 C>A

Genotyping for -163 C>A (rs762551) was performed using TaqMan technology (Applied Biosystems (ABI), Warrington UK). A 152 bp region containing the -163 C>A SNP was

amplified using the primers *CYP1A2*-1F-M163F CCAGCGTTCATGTTGGGAATCT and *CYP1A2*-1F-M163R ACTGATGCGTGTTCTGTGCTT. The fluorogenic probes VIC CGTCCTGGGCCCAC and FAM CGTCCTGTGCCCAC were included in the PCR to detect the presence of a G (C on forward strand) or T (A on forward strand) at position -163 respectively. The design of the probes was such that if they annealed specifically between the forward and reverse primers, the sequence specific signal was generated as a direct result of probe degradation during PCR by the 5' nuclease activity of the polymerase.

DNA was amplified in 384 well microplates and in 4 µl reaction volumes containing 1 µl of 1 ng/µl DNA, 2 µl of 1x TaqMan Genotyping Master Mix (Applied Biosystems (ABI), Warrington UK), 0.05 µl of 80x assay mix (containing primers and probes from Applied Biosystems (ABI), Warrington UK) and 0.95 µl of sterile water. The thermal cycler conditions were: 10 minutes of pre-incubation at 95 ºC, followed by 40 cycles of 15 seconds at 92 ºC and 1 minute at 60 ºC. The resultant PCR product was analysed using TaqMan 7900HT software (Applied Biosystems (ABI), Warrington UK).

2.2.9 Genotyping of 2159 G>A

Genotyping of 2159 G>A (rs2472304) was performed using the ABI TaqMan SNP genotyping assay C__11772996_1_ (Applied Biosystems (ABI), Warrington UK). This assay contained a mix of unlabeled PCR primers and TaqMan MGB probes (FAM and VIC dye-labeled) for the allelic discrimination of rs2472304. Details of the oligonucleotides are not disclosed and are patented by ABI (US patents and corresponding patent claims outside the US: 5,538,848, 5,723,591, 5,876,930, 6,030,787, 6,258,569, and 5,804,375 (claims 1-12 only)). The amplicon sequence is however made available by ABI to reviewers on request.

DNA was amplified in 384 well microplates and in 4 µl reaction volumes containing 1 µl of 1 ng/µl DNA, 2 µl of 1x TaqMan Genotyping Master Mix (Applied Biosystems (ABI), Warrington UK), 0.2 µl of 20x assay mix (containing primers and probes from Applied Biosystems (ABI), Warrington UK) and 0.8 µl of sterile water. The thermal cycler conditions were: 10 minutes of pre-incubation at 95 ºC, followed by 40 cycles of 15 seconds at 92 ºC and 1 minute at 60 ºC. The resultant PCR product was analysed using TaqMan 7900HT software (Applied Biosystems (ABI), Warrington UK).

2.2.10 Statistical analysis

2.2.10.1 Hardy-Weinberg equilibrium (HWE)

HWE defines the expected proportions of genotypes with respect to observed allele frequencies in a randomly mating population, and is given by the formula: $p^2 + 2pq + q^2 = 1$, where p is the major allele frequency and $q = 1 - p$. Deviations from HWE can be observed due to non-random

mating (e.g. inbreeding and assortive mating), selection, population stratification, or practical issues (e.g. allele drop out and small sample size). Tests for departure of observed genotype frequencies from those expected under HWE were performed using Arlequin software (Schneider et al., 2000). Arlequin uses a Fisher's exact test analogue (Guo and Thompson, 1992) to evaluate departures from HWE.

2.2.10.2 Fisher's exact test

A Fisher's exact test was used to test for the significance of associations between two categorical variables in a 2 x 2 contingency table (e.g. number of mutated sites versus number of unmutated sites in exons 2 and 3). Under the null hypothesis, there is no association between row and column classifications. A 2 x 2 table of expected cell frequencies, under the null hypothesis, with the same row and column totals as the observed is generated. The test calculates the difference between the data observed and the data expected under the null hypothesis. The probability for the test is calculated by generating all tables that are similar to, or more extreme than the observed table. The p values of these tables and the p value of the observed table are summed to give the exact p value for the test. Pairwise Fisher's exact tests were performed using GraphPad InStat version 3.00 for Windows 95, GraphPad Software, San Diego California USA, www.graphpad.com.

2.2.10.3 Pairwise linkage disequilibrium (LD)

Pairwise LD (the non-random association of alleles at two different bi-allelic loci) was measured, using the D' parameter (Lewontin, 1964), using GOLD software (Abecasis and Cookson, 2000). D is calculated according to: $D = P_{AB} - (P_A \times P_B)$, where P_{AB} is the observed frequency of haplotype AB, and $P_A \times P_B$ is the expected frequency of haplotype AB under linkage equilibrium (i.e. the product of the observed allele frequencies of A and B). Since D is dependent on the frequencies of alleles, it is not always comparable between loci. To improve comparability, D is normalised to give the parameter D'. The absolute value of D' is calculated by dividing D by its maximum possible value given the allele frequencies of the two loci. A chi-square test is also performed to determine whether the observed haplotype distribution is significantly different from the expected haplotype distribution under linkage equilibrium.

2.2.10.4 Haplotype inference

Haplotype phase inference was estimated from unphased population genotype data using five different approaches: an approach implemented in the programme fastPHASE (Scheet and Stephens, 2006), the maximum-likelihood approach implemented through the EM algorithm (Excoffier and Slatkin, 1995) in Arlequin software, the ELB approach (Excoffier et al., 2003) also

implemented in Arlequin software and two manual approaches based on Clark's algorithm (Clark, 1990). One approach assigns precedence (designated choice) in inferring phase in heterozygotes to the ancestral haplotype and is herein called 'Clark-ancestral'. The other approach assigns precedence (designated choice) to the most common possible haplotype identified from homozygotes and is herein named 'Clark-common'. Attempts were made to infer haplotypes using a Bayesian statistical approach implemented through Phase 2.0 software (Stephens et al., 2001), however Phase was unable to handle the quantity of data. Details of all haplotype estimation approaches are described in appendix 1.

Haplotype counts from the different inference methods were compared using Pearson's correlation coefficient analyses in Microsoft Office Excel 2003. Pearson's correlation (r) reflects the degree of the linear relationship between two variables (i.e. haplotype counts from two haplotype inference methods). Pearson's r ranges from +1 to -1. A perfect positive linear relationship between variables is given by +1, a perfect negative linear relationship is given by -1. When r = 0, the two variables are not correlated.

2.2.10.5 Gene diversity **

Gene diversity (h) (probability of randomly choosing two different haplotypes from a sample (equivalent to heterozygosity for diploid data)) was estimated according to unbiased formulae of Nei (1987): $H = n(1 - \sum_{xi}^{2})/(n - 1)$, where n is the number of gene copies, $_{xi}$ is the frequency of the $_{i}$th allele. Standard deviation of H was the square root of the variance of H (Nei, 1987).

2.2.10.6 Nucleotide diversity **

Nucleotide diversity (π) (probability that two randomly chosen homologous nucleotides are different (equivalent to gene diversity at the nucleotide level)) was estimated according to Nei (1987): $\pi = n(\sum x_i x_j \pi_{ij})/(n-1)$, where n is the number of sequences, x_i and x_j the frequencies of the ith and jth sequences respectively and π_{ij} the proportion of different nucleotides between them. Standard deviation of π was the square root of the variance of π (Nei, 1987).

2.2.10.7 Exact test of pairwise population differentiation **

Genetic differences between populations were assessed using an Exact test of pairwise population differentiation with 10,000 Markov steps (Rousset and Raymond, 1995; Goudet et al., 1996). This test is analogous to a Fisher's Exact test (Lee et al., 2004) but the size of the 2 x 2 contingency table is extended to the number of populations being compared (two in a pairwise population comparison, two or greater in a global test) by the total number of different haplotypes present. All potential states of the contingency table are explored with a random

walk via a Markov chain model and estimates of the probability of observing a table less or equally likely than the observed data under the null hypothesis of panmixia are obtained. Populations are considered to be significantly different if the *p* value is smaller than the significance level (set at 5 %).

2.2.10.8 Genetic distance (F_{ST}) **

F_{ST} measures the genetic variability within and between populations, comparing mean genetic diversity within sub-populations to genetic diversity in the meta-population, as follows: $F_{ST} = (H_T - H_S)/H_{T,}$ where H_T is the expected heterozygosity of the meta-population and H_S is the mean expected heterozygosity across sub-populations (Hudson et al., 1992). Genetic distance between two populations was analysed using population pairwise F_{ST} values (Reynolds et al., 1983), whilst the apportionment of diversity within and between more than two populations was analysed using hierachical F_{ST} values (Excoffier et al., 1992).

All statistically analyses denoted with ** were performed using Arlequin software (Schneider et al., 2000) and hypotheses were tested using permutation analysis (p values were the proportion of permutations giving, for example, an F_{ST} value ≥ the value obtained for the observed data), thus the data was not assumed to have a normal distribution.

2.2.10.9 Principal coordinates analysis

Principal coordinates analysis (Gower, 1966) was performed, using the R statistical package (www.R-project.org), on pairwise similarity matrices. Similarity was quantified as being equal to the value of the genetic distance subtracted from 1 (1-Fst for example). Values along the main diagonal, representing the similarity of each population sample to itself, were calculated from the estimated genetic distance between two copies of the same sample. For Fst distances, the resulting similarity of a sample to itself simplifies to n/(n–1), where n = number of chromosomes.

2.2.11 Prediction of functional effect of non-synonymous changes

Effects of amino acid substitutions on the structure and function of CYP1A2 were predicted using PolyPhen software (http://genetics.bwh.harvard.edu/pph/). PolyPhen aligns human amino acid sequences against orthologous amino acid sequences, and predicts whether amino acid changes in humans will have the following effects on the structure/function of the protein: benign (no damage); possibly damaging; probably damaging; presume definite damage. Predictions consider the nature of the amino acid change and the degree of protein conservation across species.

2.3 Results

2.3.1 Summary of variation found in Ethiopian ascertainment and NIEHS populations

- In total, 49 *CYP1A2* variants were found in the Ethiopian ascertainment samples (table 2.8a) whilst 22 variants were reported in the corresponding *CYP1A2* sequence in the NIEHS samples (table 2.8d).

- All variants, except one (a single nucleotide deletion in the 3' UTR reported in both Ethiopians and NIEHS populations) were single nucleotide substitutions which together exhibited a transition to transversion ratio estimate of 23:25 in Ethiopians, compared to 5:2 in the NIEHS populations. Consequently, more than double the number of transitions than transversions were observed in the NIEHS sample set, whilst in Ethiopia there were slightly more transversions than transitions, but this difference was not significant (Fisher's exact test, $p > 0.1$).

- The ratio of coding to non-coding variant sites was 5:19 in Ethiopians and 4:7 in the NIEHS populations (Fisher's exact test, $p > 0.1$).

- Exons 1, 4 and 5 were monomorphic in both Ethiopians and the NIEHS populations. In coding exons, all substitutions except one were non-synonymous for both Ethiopians and NIEHS populations (a synonymous to non-synonymous ratio estimate of 1:9 and 1:7 was observed in the Ethiopians and the NIEHS populations respectively).

- Notably, significantly more non-synonymous mutations were found in exon 7 compared to all other exons collectively, in Ethiopians (Fisher's exact test comparing numbers of non-synonymous mutation sites against combined numbers of not mutated and synonymous mutation sites, in exon 7 versus other exons collectively, $p < 0.02$).

- Of the 49 variants detected in Ethiopians, 19 have previously been reported (table 2.8b), while 30 were novel (table 2.8c).

- In both Ethiopians and NIEHS populations, variant sites were not observed within 17 bases either side of each intron/exon boundary and all reported catalytic residues (amino acids D320 and T321 in exon 4, and F451 and C458 in exon 7) were monomorphic.

2.3.2 Frequencies of *CYP1A2* variants

CYP1A2 SNP frequencies are shown in table 2.9. No SNP frequency for any population deviated significantly from HWE at the 1 % significance level. Only three SNP loci were variable

Table 2.8 *CYP1A2* variants a) observed in, b) confirmed by, c) added by the Ethiopian ascertainment samples and d) observed in the NIEHS populations

Location	Substitutions				Transitions				Transversions				Insertion/Deletions				Short tandem repeat				Splice site				Synonymous				Non synonymous			
	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
5' end	2	0	2	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0												
3' end	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0												
3' UTR	12	5	7	4	5	3	2	3	7	2	5	1	1	1	0	1	0	0	0	0												
Intron 1	7	4	3	4	3	2	1	3	4	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0								
Intron 2	3	1	2	1	3	1	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0								
Intron 3	2	0	2	2	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0								
Intron 4	2	2	0	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0								
Intron 5	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0								
Intron 6	9	3	6	1	4	3	1	1	5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0								
Exon 1 (5' UTR)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0												
Exon 2	1	1	0	4	0	0	0	3	1	1	0	1	0	0	0	0	0	0	0	0					0	0	0	0	1	1	0	4
Exon 3	1	1	0	2	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0					0	0	0	0	1	1	0	2
Exon 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					0	0	0	0	0	0	0	0
Exon 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					0	0	0	0	0	0	0	0
Exon 6	2	0	2	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0					0	0	0	0	2	0	2	0
Exon 7	6	1	5	2	4	1	3	2	2	0	2	0	0	0	0	0	0	0	0	0					1	1	0	1	5	0	5	1

Table 2.9 *CYP1A2* allele frequencies in the Ethiopian ascertainment and NIEHS populations

Yellow = novel mutations, green = known mutations, *italics* = private to one population

f = frequency, n = chromosome number

CYP1A2 variant Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1)	NCBI dbSNP database refSNP ID(s)	Chromosome position in human reference assembly 36	Location	Amino acid change	Afar		Amhara		Anuak		Maale		Oromo		African American		Yoruba		European		Hispanic		East Asian	
					f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n
-1014 C>A		*72828121*	*5' upstream*		*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
-1008 G>A		*72828127*	*5' upstream*		*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
-739 T>G	rs2069526	72828396	Intron 1		0.05	8	0.07	11	0.03	5	0.08	12	0.09	14	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
-729 C>T	rs12720461	72828406	Intron 1		0.00	0	0.01	1	0.00	0	0.01	2	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
-661 T>A		*72828474*	*Intron 1*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
-592 C>T		*72828543*	*Intron 1*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
-569 G>A	*rs45518531*	*72828566*	*Intron 1*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.03*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
-505 G>A	rs45607039	72828630	Intron 1		0.00	0	0.00	0	0.07	11	0.00	0	0.00	0	0.03	1	0.04	1	0.00	0	0.00	0	0.00	0
-163 C>A	rs762551	72828972	Intron 1		0.57	86	0.58	89	0.38	57	0.43	65	0.62	94	0.65	17	0.58	14	0.66	29	0.86	36	0.63	30
-151 G>T		72828984	Intron 1		0.00	0	0.01	2	0.00	0	0.01	2	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
-61 A>G	*rs41279194*	*72829074*	*Intron 1*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.02*	*1*	*0.00*	*0*	*0.00*	*0*
53 C>G	rs17861152	72829187	Exon 2	S18C	0.00	0	0.01	1	0.01	1	0.00	0	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
217 G>A	*rs45565238*	*72829351*	*Exon 2*	*G73R*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.04*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
310 G>A	*rs34067076*	*72829444*	*Exon 2*	*D104N*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.02*	*1*	*0.00*	*0*	*0.00*	*0*
331 C>T	*rs45442197*	*72829465*	*Exon 2*	*L111F*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.03*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
613 T>G	*rs45540640*	*72829747*	*Exon 2*	*F205V*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.04*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
869 G>C	*rs45533242*	*72830003*	*Intron 2*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.02*	*1*	*0.00*	*0*
1352 G>A	rs34264399	72830486	Intron 2		0.01	2	0.01	1	0.00	0	0.02	3	0.01	2	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
1353 G>A		*72830487*	*Intron 2*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
1370 G>A		*72830504*	*Intron 2*		*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
1460 C>T	*rs45468096*	*72830594*	*Exon 3*	*R281W*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.02*	*1*	*0.00*	*0*	*0.00*	*0*
1513 C>A	rs17861157	72830647	Exon 3	S298R	0.03	4	0.01	1	0.11	16	0.05	8	0.05	7	0.03	1	0.13	3	0.00	0	0.00	0	0.00	0
1589 G>T		72830723	Intron 3		0.01	1	0.04	6	0.00	0	0.06	9	0.05	7	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
1611 G>A		72830745	Intron 3		0.00	0	0.01	2	0.00	0	0.00	0	0.02	3	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
1649 G>T	*rs45484991*	*72830783*	*Intron 3*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.02*	*1*
1669 C>T	*rs45445793*	*72830803*	*Intron 3*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.04*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
2159 G>A	rs2472304	72831293	Intron 4		0.38	58	0.43	66	0.07	10	0.23	35	0.36	54	0.13	4	0.00	0	0.63	25	0.33	13	0.13	6
2321 G>C	rs3743484	72831455	Intron 4		0.03	4	0.02	3	0.02	3	0.01	2	0.02	3	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
2534 C>T		*72831668*	*Intron 5*		*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
3463 C>T		72832597	Exon 6	T395M	0.01	2	0.01	1	0.01	2	0.01	1	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
3468 A>C		*72832602*	*Exon 6*	*N397H*	*0.01*	*2*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
3588 G>T		*72832722*	*Intron 6*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
3605 A>G	rs34356615	72832739	Intron 6		0.01	1	0.00	0	0.00	0	0.00	0	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
3613 T>C	rs4646427	72832747	Intron 6		0.06	9	0.07	11	0.03	5	0.07	11	0.09	13	0.11	3	0.08	2	0.00	0	0.02	1	0.02	1
4957 C>G		72834091	Intron 6		0.00	0	0.01	2	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
4961 C>T		*72834095*	*Intron 6*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
4984 C>G		*72834118*	*Intron 6*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5010 C>T	*rs28399423*	*72834144*	*Intron 6*		*0.01*	*2*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5015 C>G		*72834149*	*Intron 6*		*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5029 C>G		*72834163*	*Intron 6*		*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5094 T>C		72834228	Exon 7	F432S	0.00	0	0.01	1	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
5105 G>A		72834239	Exon 7	D436N	0.01	1	0.01	1	0.01	1	0.07	10	0.02	3	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
5112 C>T	*rs45486893*	*72834246*	*Exon 7*	*T438I*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.03*	*1*
5253 C>G		*72834387*	*Exon 7*	*P485R*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5284 C>A		*72834418*	*Exon 7*	*Y495Ter*	*0.00*	*0*	*0.00*	*0*	*0.03*	*4*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5328 G>A		72834462	Exon 7	R510Q	0.01	2	0.00	0	0.00	0	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
5347 C>T	rs2470890	72834481	Exon 7	Synonymous	0.29	44	0.33	51	0.04	6	0.13	20	0.25	38	0.13	4	0.00	0	0.63	25	0.33	12	0.12	6
5355 G>C		*72834489*	*3' UTR*		*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
5521 A>G	rs11636419	72834655	3' UTR		0.10	15	0.09	13	0.07	10	0.09	14	0.11	16	0.17	5	0.09	2	0.03	1	0.08	3	0.23	11
5620 A>C	rs58661304	72834754	3' UTR		0.05	8	0.03	4	0.12	17	0.20	30	0.06	9	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
5987 G>T		72835121	3' UTR		0.01	1	0.00	0	0.00	0	0.04	6	0.02	3	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
6021 C>T	*rs45485796*	*72835155*	*3' UTR*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.02*	*1*	*0.00*	*0*
6233 C>T		*72835367*	*3' UTR*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*
6324 Gdel	rs45564134	72835458	3' UTR		0.04	6	0.01	2	0.06	9	0.00	0	0.02	2	0.17	5	0.13	3	0.00	0	0.00	0	0.00	0
6388 T>G		72835522	3' UTR		0.01	1	0.00	0	0.00	0	0.01	1	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
6537 G>A	rs45599945	72835671	3' UTR		0.00	0	0.00	0	0.03	4	0.00	0	0.00	0	0.00	0	0.04	1	0.00	0	0.00	0	0.00	0
6562 T>A		72835696	3' UTR		0.01	1	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
6674 C>G	rs17861162	72835808	3' UTR		0.08	12	0.09	14	0.09	14	0.09	14	0.11	16	0.19	5	0.15	3	0.03	1	0.08	3	0.22	8
6685 A>G	rs17861163	72835819	3' UTR		0.03	5	0.01	1	0.00	0	0.00	0	0.02	3	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
6714 A>T		72835848	3' UTR		0.00	0	0.01	1	0.00	0	0.00	0	0.01	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
6765 C>T		*72835899*	*3' UTR*		*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.01*	*1*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*	*0.00*	*0*

in all populations and these were: -163 C>A (intron 1), 5521 A>G and 6674 C>G (both in the 3' UTR). Half of the SNPs were private to populations (highlighted in italics in table 2.9). Notably, all variants found more than once in any one NIEHS population, were detected in the Ethiopian ascertainment population. It is also relevant to note that no non-synonymous mutation exceeded 13 % in any one population with many being observed at frequencies of 1 - 3 %. In addition, a previously unreported premature stop codon in exon 7 (5384 C>A resulting in Y495Ter) was observed in Anuak at 3 %.

2.3.3 LD across *CYP1A2*

Pairwise LD (D´) was measured across *CYP1A2* both by pooling all the Ethiopian ascertainment and NIEHS populations data together (world dataset) and by analysing each group separately. Monomorphic loci and rare variants (where frequency < 0.01) were removed from the datasets prior to the analysis. In the world dataset (figure 2.3), the majority of *CYP1A2* loci are in total LD (D´ = 1) but several cases where D´ was less than 1 were observed across the gene. The majority of lower D´ values were evident between pairs of loci including at least one marker towards the 3' end of the gene, and loci from intron 1 up to and including 5521 in the 3' UTR constituted an LD block as defined by other investigators (figure 2.3). When independently analysing pairwise LD in each population, D´ values of less than 1 were observed in all groups except Yoruba, Europeans and East Asians (figure 2.4).

2.3.4 *CYP1A2* haplotype inference

Given that rare *CYP1A2* variants were observed in this study, and that *CYP1A2* could not be analysed as one LD block (figure 2.3), haplotype inference for the entire *CYP1A2* gene (61 variants) was not straight forward. *CYP1A2* haplotype estimates were compared from the following haplotype inference approaches: ELB (Excoffier et al., 2003), EM (Excoffier and Slatkin, 1995), fastPHASE (Scheet and Stephens, 2006) and two extensions of the Clark algorithm (Clark, 1990), Clark-ancestral and Clark-common, which were performed by hand. Attempts were made to estimate haplotypes via Phase (Stephens et al., 2001) but the software was unable to handle the large quantity of data. The Ethiopian ascertainment and NIEHS population genotype data were pooled (world dataset) and haplotype inference was performed using this dataset in the first instance. Haplotypes were then inferred from the pooled Ethiopian ascertainment dataset and from each of the individual populations. Note that only samples with full genotype data were included in the analysis.

Pearson's correlation coefficient analysis was performed to compare the haplotype counts generated from the different haplotype inference approaches. Haplotype counts and Pearson's correlation coefficients are shown in supplementary tables S1-S12 (please see CD). In summary, a strong positive correlation ($r \geq 0.8$) was observed among all approaches, except

Figure 2.3 Pairwise LD (D') across *CYP1A2* in the world dataset. *CYP1A2* variants and their relative locations within the gene are highlighted in grey, D' values of 1 are highlighted in pink, significant Chi square associations are in bold ($p < 0.05$). The area bordered in red constitutes an LD block as defined by Gabriel et al. (2002) in Haploview (www.hapmap.org/haploview).

Intron 1	-739															
	-505	1.00														
	-163	**1.00**	**1.00**													
Exon 3	1513	0.99	0.08	**1.00**												
Intron 3	1589	**1.00**	1.00	**1.00**	0.96											
Intron 4	2159	**1.00**	**1.00**	**0.99**	**1.00**	0.98										
	2321	0.75	1.00	**1.00**	1.00	1.00	1.00									
Intron 6	3613	**0.94**	0.75	**0.93**	1.00	**1.00**	**0.85**	0.84								
Exon 7	5105	1.00	1.00	**1.00**	1.00	1.00	0.79	1.00	1.00							
	5347	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**0.99**	0.68	**1.00**	1.00						
3' UTR	5521	**0.98**	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**	**0.93**	1.00	**1.00**					
	5620	0.95	0.10	0.26	0.23	0.82	**0.79**	0.05	**1.00**	0.06	**1.00**	**1.00**				
	5987	0.12	1.00	**1.00**	**1.00**	0.05	0.89	1.00	0.11	1.00	0.28	0.85	0.56			
	6324	0.43	0.10	**1.00**	1.00	0.52	**1.00**	0.02	0.63	1.00	**1.00**	0.85	**0.65**	1.00		
	6674	**0.92**	1.00	**0.95**	1.00	**0.89**	**1.00**	**1.00**	**0.91**	1.00	**1.00**	**0.96**	0.94	0.00	0.89	
		-739	-505	-163	1513	1589	2159	2321	3613	5105	5347	5521	5620	5987	6324	6674
		Intron 1			Exon 3	Intron 3	Intron 4		Intron 6	Exon 7		3' UTR				

Figure 2.4 Pair wise LD (D´) across *CYP1A2* in the various populations. *CYP1A2* variants and their relative locations within the gene are highlighted in grey, D' values of 1 are highlighted in pink, significant Chi square associations are in bold ($p < 0.05$).

African American

		-163	2159	3613	5347	5521	6324	6674
		Intron 1	Intron 4	Intron 6	Exon 7	3' UTR		
Intron 1	-163							
Intron 4	2159	0.30						
Intron 6	3613	0.28	1.00					
Exon 7	5347	0.30	**1.00**	1.00				
3' UTR	5521	1.00	1.00	**1.00**	1.00			
	6324	1.00	0.83	1.00	0.83	1.00		
	6674	1.00	1.00	**1.00**	1.00	**1.00**	1.00	

Hispanic

		-163	2159	5347	5521	6674
		Intron 1	Intron 4	Exon 7	3' UTR	
Intron 1	-163					
Intron 4	2159	1.00				
Exon 7	5347	1.00	**1.00**			
3' UTR	5521	0.57	1.00	1.00		
	6674	0.59	1.00	1.00	**1.00**	

European

		-163	2159	5347
		Intron 1	Intron 4	Exon 7
Intron 1	-163			
Intron 4	2159	**1.00**		
Exon 7	5347	**1.00**	**1.00**	

Yoruba

		-163	1513	3613	5521	6324	6674
		Intron 1	Exon 3	Intron 6	3' UTR		
Intron 1	-163						
Exon 3	1513	1.00					
Intron 6	3613	1.00	1.00				
3' UTR	5521	1.00	1.00	**1.00**			
	6324	1.00	1.00	1.00	1.00		
	6674	1.00	1.00	**1.00**	**1.00**	1.00	

East Asian

		-163	2159	5347	5521	6674
		Intron 1	Intron 4	Exon 7	3' UTR	
Intron 1	-163					
Intron 4	2159	1.00				
Exon 7	5347	1.00	**1.00**			
3' UTR	5521	**1.00**	1.00	1.00		
	6674	**1.00**	1.00	1.00	**1.00**	

Afar

		-739	-163	1352	1513	2159	2321	3463	3468	3613	5010	5328	5347	5521	5620	6324	6674	6685
		Intron 1		Intron 2	Exon 3	Intron 4		Exon 6		Intron 6		Exon 7		3' UTR				
Intron 1	-739																	
	-163	1.00																
Intron 2	1352	1.00	1.00															
Exon 3	1513	1.00	1.00	1.00														
Intron 4	2159	**1.00**	**0.95**	1.00	1.00													
	2321	1.00	0.99	1.00	1.00	1.00												
Exon 6	3463	1.00	1.00	1.00	1.00	**1.00**	1.00											
	3468	1.00	1.00	1.00	1.00	1.00	1.00	1.00										
Intron 6	3613	**0.87**	0.46	1.00	1.00	0.66	1.00	1.00	1.00									
	5010	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00								
Exon 7	5328	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00							
	5347	1.00	**1.00**	0.45	0.47	**0.96**	0.47	0.46	1.00	**1.00**	1.00	1.00						
3' UTR	5521	**1.00**	**1.00**	1.00	1.00	**1.00**	**0.69**	1.00	1.00	**0.74**	1.00	1.00	**1.00**					
	5620	1.00	0.30	1.00	0.16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
	6324	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	**1.00**			
	6674	**0.84**	**1.00**	1.00	1.00	**1.00**	**1.00**	1.00	1.00	**0.67**	1.00	1.00	1.00	**1.00**	1.00	1.00		
	6685	1.00	1.00	1.00	1.00	1.00	0.29	1.00	1.00	1.00	1.00	1.00	1.00	0.02	0.10	0.13	0.04	

Amhara

		-739	-163	-151	1589	1611	2159	2321	3613	4957	5347	5521	5620	6324	6674
		Intron 1			Intron 3		Intron 4		Intron 6		Exon 7	3' UTR			
Intron 1	-739														
	-163	**1.00**													
	-151	1.00	1.00												
Intron 3	1589	**1.00**	**1.00**	1.00											
	1611	**1.00**	1.00	1.00	1.00										
Intron 4	2159	**1.00**	**1.00**	1.00	1.00	1.00									
	2321	0.22	1.00	1.00	1.00	**0.48**	1.00								
Intron 6	3613	**1.00**	**1.00**	1.00	**1.00**	1.00	**1.00**	0.22							
	4957	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00						
Exon 7	5347	**1.00**	**1.00**	1.00	1.00	0.95	**1.00**	0.01	**1.00**	0.89					
3' UTR	5521	**0.90**	**1.00**	1.00	**1.00**	1.00	**1.00**	**1.00**	**0.90**	1.00	1.00				
	5620	0.12	1.00	**1.00**	0.19	1.00	1.00	1.00	0.12	1.00	0.95	0.10			
	6324	0.45	1.00	1.00	0.46	1.00	1.00	1.00	0.44	1.00	1.00	0.44	**1.00**		
	6674	**1.00**	**1.00**	1.00	**1.00**	1.00	**1.00**	1.00	**1.00**	1.00	**1.00**	**1.00**	0.07	0.43	

Anuak

		-739	-505	-163	1513	2159	2321	3463	3613	5284	5347	5521	5620	6324	6537	6674
		Intron 1			Exon 3	Intron 4		Exon 6	Intron 6	Exon 7		3' UTR				
Intron 1	-739															
	-505	1.00														
	-163	**1.00**	**1.00**													
Exon 3	1513	1.00	1.00	**1.00**												
Intron 4	2159	0.08	1.00	**1.00**	1.00											
	2321	1.00	1.00	1.00	1.00	1.00										
Exon 6	3463	1.00	1.00	1.00	1.00	**1.00**	1.00									
Intron 6	3613	**1.00**	1.00	**1.00**	1.00	0.08	1.00	1.00								
Exon 7	5284	1.00	1.00	**1.00**	**1.00**	1.00	1.00	1.00	1.00							
	5347	0.14	1.00	**1.00**	0.48	**1.00**	1.00	1.00	0.14	1.00						
3' UTR	5521	**1.00**	1.00	**1.00**	1.00	0.62	**0.62**	1.00	**1.00**	1.00	0.04					
	5620	0.33	0.06	**0.75**	1.00	0.05	0.57	1.00	0.33	0.02	0.67	0.10				
	6324	0.09	0.09	**1.00**	1.00	0.16	0.25	1.00	0.09	0.15	1.00	0.10	**1.00**			
	6537	1.00	1.00	0.33	0.17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
	6674	**1.00**	1.00	**0.84**	1.00	0.99	**1.00**	1.00	**1.00**	1.00	0.21	**1.00**	0.11	0.05	1.00	

Maale

		-739	-729	-163	-151	1352	1513	1589	2159	2321	3613	5105	5347	5521	5620	5987	6674
		Intron 1				Intron 2	Exon 3	Intron 3	Intron 4		Intron 6	Exon 7		3' UTR			
Intron 1	-739																
	-729	**1.00**															
	-163	**1.00**	1.00														
	-151	1.00	1.00	1.00													
Intron 2	1352	1.00	1.00	1.00	1.00												
Exon 3	1513	0.48	1.00	**1.00**	1.00	1.00											
Intron 3	1589	**1.00**	1.00	**1.00**	1.00	1.00	0.01										
Intron 4	2159	0.32	0.07	**1.00**	1.00	1.00	0.58	0.88									
	2321	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07								
Intron 6	3613	**1.00**	**1.00**	**1.00**	1.00	1.00	0.33	**1.00**	0.66	1.00							
Exon 7	5105	1.00	1.00	**1.00**	1.00	0.24	1.00	1.00	0.42	1.00	1.00						
	5347	0.95	0.33	**1.00**	1.00	1.00	0.16	1.00	**1.00**	1.00	0.73	0.62					
3' UTR	5521	**1.00**	**1.00**	**1.00**	1.00	1.00	0.77	**1.00**	0.18	**1.00**	**1.00**	1.00	1.00				
	5620	**1.00**	1.00	**0.75**	**1.00**	0.16	0.99	1.00	**1.00**	1.00	**1.00**	0.53	**1.00**	**1.00**			
	5987	0.03	1.00	**1.00**	1.00	1.00	**1.00**	0.05	0.86	1.00	0.03	1.00	0.09	0.05	1.00		
	6674	**1.00**	1.00	**1.00**	1.00	1.00	0.08	**1.00**	1.00	**1.00**	**1.00**	1.00	1.00	**1.00**	**1.00**	0.35	

Oromo

		-739	-163	1352	1513	1589	1611	2159	2321	3613	5105	5347	5521	5620	5987	6324	6674	6685
		Intron1		Intron 2	Exon 3	Intron 3		Intron 4		Intron 6	Exon 7		3' UTR					
Intron1	-739																	
	-163	**1.00**																
Intron 2	1352	1.00	1.00															
Exon 3	1513	0.51	1.00	1.00														
Intron 3	1589	**1.00**	1.00	1.00	1.00													
	1611	**1.00**	1.00	1.00	0.27	1.00												
Intron 4	2159	**1.00**	**1.00**	0.99	1.00	1.00	1.00											
	2321	1.00	0.68	**0.48**	1.00	1.00	1.00	1.00										
Intron 6	3613	**1.00**	1.00	1.00	0.46	**1.00**	**1.00**	**1.00**	1.00									
Exon 7	5105	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00								
	5347	**1.00**	**1.00**	0.00	0.46	1.00	1.00	**0.95**	0.96	1.00	0.96							
3' UTR	5521	**1.00**	**1.00**	0.36	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	**1.00**						
	5620	0.83	0.18	1.00	1.00	1.00	1.00	0.22	1.00	1.00	1.00	1.00	1.00					
	5987	0.16	1.00	1.00	**1.00**	1.00	0.31	1.00	1.00	0.17	1.00	0.96	1.00	1.00				
	6324	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**	1.00			
	6674	**0.82**	**1.00**	1.00	0.85	**0.65**	**1.00**	**1.00**	**1.00**	**0.80**	1.00	**1.00**	**1.00**	0.65	0.11	1.00		
	6685	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.66**	0.96	1.00	1.00	1.00	1.00	1.00	

Clark-ancestral, for haplotype counts from all datasets and the fastPHASE and EM approaches were often in complete agreement (r = 1). The majority of the weaker haplotype correlations stemmed from the Clark-common approach. These were however likely to be due to the method not resolving rare variants into haplotypes and producing more orphan alleles (unresolved haplotypes) compared to other approaches. Haplotype counts from the Clark-ancestral approach showed the weakest correlations of all and Pearson's correlation coefficients dropped to 0.2 in some comparisons in African Americans and Yoruba (supplementary tables S8 and S9 respectively). Weak correlations such as these are likely to result from differences (in haplotype counts) being inflated in and information (from which haplotypes are resolved) being lost in small datasets. Andres et al. (2007) demonstrated that, in populations of mixed ancestry, the most accurate haplotypes are probably resolved from the largest pooled sample, despite theoretical problems associated with pooling across heterogeneous population samples. In this study, a strong positive correlation ($r \geq 0.94$) was observed among haplotype counts from all approaches, including the Clark-ancestral approach, when the world dataset was analysed (supplementary table S1). In light of these findings, subsequent analyses in this study used haplotypes (for the entire *CYP1A2* gene) estimated from the world dataset and from one haplotype inference approach. The ELB approach was chosen because it is particularly well suited to problems involving many loci and/or relatively large genomic regions, including those with variable recombination rates (Excoffier et al., 2003).

2.3.5 *CYP1A2* haplotypes in the Ethiopians and NIEHS populations

2.3.5.1 *CYP1A2* (entire gene) haplotypes

A total of 64 *CYP1A2* (entire gene) haplotypes were determined/inferred in the world dataset (table 2.10). Haplotypes 1, 5 and 61 are *CYP1A2**1B, *1M and *1F respectively, which are associated with an unknown CYP1A2 activity (table 2.6). It should be noted however, that the enhancer region (thought to be -3990 to -2118 nucleotides from the translational start site (Soyama et al., 2005)) was not sequenced in the Ethiopians nor the NIEHS populations and the following SNPs, reported in a number of *CYP1A2** alleles by the Human CYP450 Nomenclature Committee, were not genotyped in this study: -3860G>A, -3594T>G, -3113A>G, -3053A>G, -2467delT, -2808A>C and -2667T>G. As a consequence of this, any *CYP1A2** haplotype assignments in this study should be considered preliminary until the enhancer region is sequenced. All haplotypes shown in table 2.10, except haplotypes 1, 5 and 61, are not reported by the Human CYP450 Nomenclature Committee, with the consequence that 61 novel haplotypes have been generated in this study. Many of these novel haplotypes are however closely related to those previously reported. For example, *CYP1A2**1K, with -729C>T, -739T>G and -163C>A, and coding for a protein with a reduced function (table 2.6), was not observed in this study. All three nucleotide changes were however identified in haplotype 49 (table 2.10) which consequently may also have a reduced function. Notably, 19 novel haplotypes, with non-synonymous mutations, were identified in this study (highlighted in pink in table 2.10).

Table 2.10 Haplotype inference across the entire *CYP1A2* gene in the Ethiopian ascertainment and NIEHS populations. Frequencies are shown in figure 2.5. Within the world dataset (352 individuals), which only included samples with full genotype data, 54 polymorphic *CYP1A2* loci and 231 ambiguous individuals (where phase is not inferred with high confidence) were observed. The ELB haplotype inference approach estimated 64 *CYP1A2* haplotypes from this sample.

Nucleotide change [1]	-1014 C>A	-1008 G>A	-739 T>G	-729 C>T	-592 C>T	-569 G>A	-505 G>A	-163 C>A	-151 G>T	-61 A>G	53 C>G	217 G>A	331 C>T	613 T>G	869 G>C	1352 G>A	1370 G>A	1513 C>A	1589 G>T	1611 G>A	1649 G>T	2159 G>A	2321 G>C	2534 C>T	3463 C>T	3468 A>C	3588 G>T	3605 A>G	3613 T>C	4957 C>G	4961 C>T	5010 C>T	5015 C>G	5029 C>G	5094 T>C	5105 G>A	5253 C>G	5284 C>A	5328 G>A	5347 T>C	5355 G>C	5521 A>G	5620 A>C	5987 G>T	6021 C>T	6233 C>T	6324 Gdel	6388 T>G	6537 G>A	6562 T>A	6674 C>G	6685 A>G	6714 A>T	6765 C>T
Location	5' upstream	5' upstream	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Exon 2	Exon 2	Exon 2	Exon 2	Intron 2	Intron 2	Intron 2	Exon 3	Intron 3	Intron 3	Intron 3	Intron 4	Intron 4	Intron 5	Exon 6	Exon 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR
Amino acid change											S18C	G73R	L111F	F205V				S298R							T395M	N397H										F432S	D436N	P485R	Y495Ter	R510Q	N516N																

Haplotype id [2]	
1	*1B
2	
3	
4	
5	*1M
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
57	
58	
59	
60	
61	*1F
62	
63	
64	

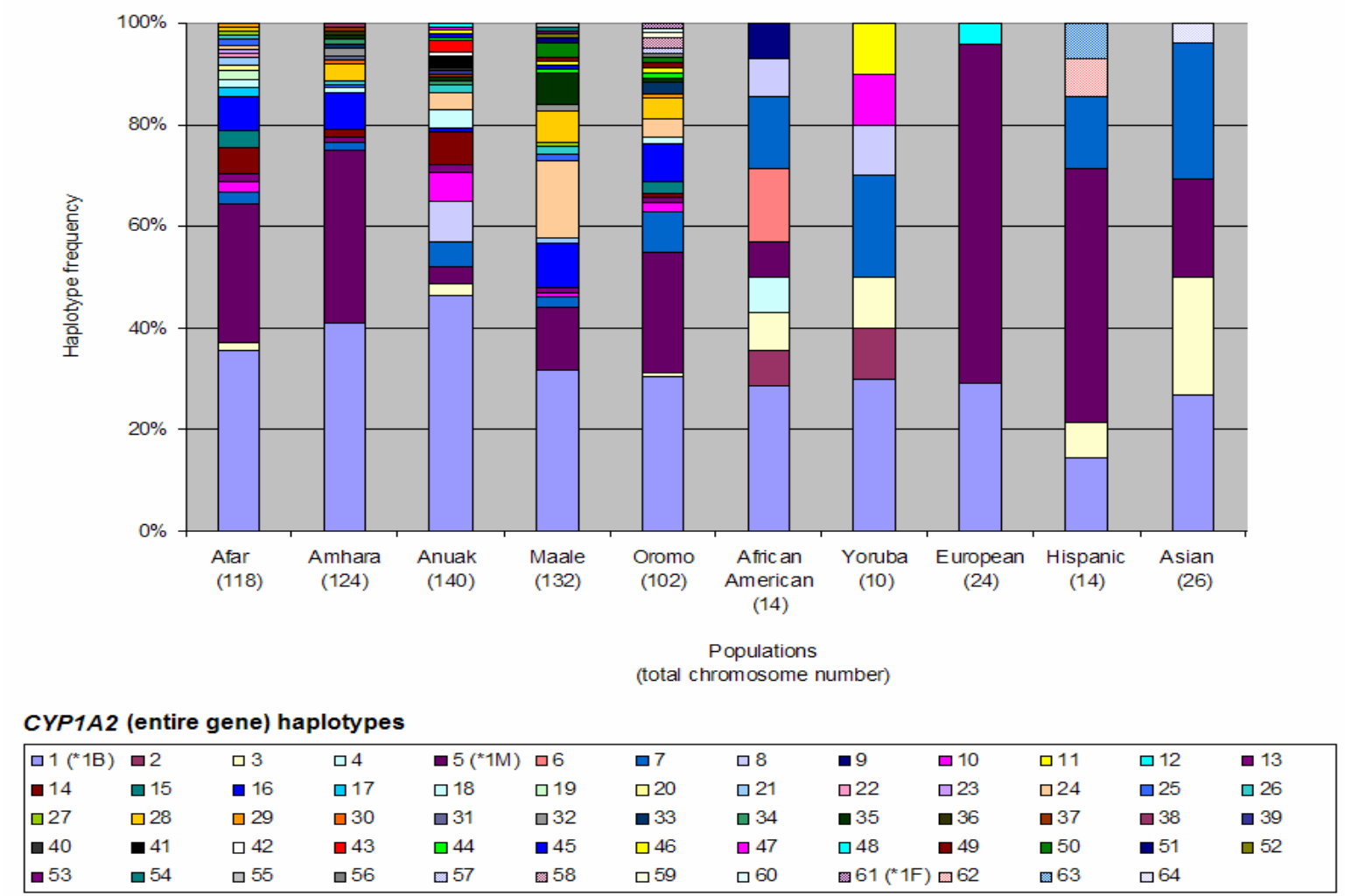
19 haplotypes containing non-synonymous mutations

[1] Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1) from the *CYP1A2* genomic reference sequence (NC_000015.8)

[2] White cell, allele observed in *CYP1A2**1A, grey cell, derived allele. A total of 21 haplotypes (shown in bold) were unambiguously resolved from homozygous genotypes at all loci or from a single site heterozygote. Depending on variation in the enhancer region (which was not sequenced in this study), haplotypes 1, 5 and 61 may be Human CYP Allele Nomenclature Committee *CYP1A2**1B, *1F and *1M respectively.

Predicted associated phenotype (refer to section 2.1.2.2): *1B, *1F and *1M = undetermined CYP1A2 activity, haplotype 49 = reduced CYP1A2 activity due to -729 C>T (Aklillu et al., 2003)

Figure 2.5 ***CYP1A2*** **(entire gene) haplotype frequencies in the Ethiopian ascertainment and NIEHS populations.** Asians are East Asians.



An extensive analysis of the cds exons is described in the next section. The evolution of *CYP1A2* in the context of the chimpanzee and other primates is discussed in chapter 3.

2.3.5.2 *CYP1A2* (entire gene) haplotype frequencies

CYP1A2 (entire gene) haplotype frequencies are shown in figure 2.5. Considerably more haplotypes were found in Ethiopians compared to NIEHS populations, however this may be a reflection of the relative sizes of the datasets (shown in brackets in figure 2.5). Haplotype 1 (*CYP1A2**1B) was the modal haplotype (≥ 30 %) in all groups with an African Ancestry whilst both haplotypes 1 and 7 were found at equal frequencies (27 %) in East Asians and haplotype 5 (*CYP1A2**1M), which was observed in all groups except Yoruba, was identified as the modal type in Europeans (67 %) and Hispanics (50 %). Haplotype 49, which is related to *CYP1A2**1K with a reduced function, was identified as singletons in Maale and Oromo, and haplotype 61 (*CYP1A2**1F) was only found as a singleton in Oromo. Notably, 49 haplotypes (13 – 61) were specific to Ethiopia while nine haplotypes (2, 4, 6, 9, 11, 12, 62 - 64) were only observed outside Ethiopia, all of which, except haplotype 6, were observed as singletons in any one group.

2.3.5.3 Recombined *CYP1A2* (entire gene) haplotypes

No recombination was observed in any of the *CYP1A2* (entire gene) haplotypes found in NIEHS populations, Afar and Amhara. Four haplotypes, and hence recombination, were however observed between pairs of SNPs in Anuak, Maale and Oromo (figure 2.6).

Figure 2.6 *CYP1A2* (entire gene) haplotype and SNP combinations from which recombination was observed in Anuak, Maale and Oromo. White cell = allele observed in CYP1A2*1A, grey cell = derived allele

Anuak

Haplotype id	-163 C>A Intron 1	6674 C>G 3' UTR
1		
7		
39		
45		

Anuak

Haplotype id	-163 C>A Intron 1	5620 A>C 3' UTR
7		
24		
41		
46		

Anuak, Maale & Oromo

Haplotype id	2159 G>A Intron 4	5620 A>C 3' UTR
7		
16		
44		
46		

Maale

Haplotype id	1513 C>A Exon 3 (S298R)	5620 A>C 3' UTR
7		
10		
46		
51		

Maale & Oromo

Haplotype id	-163 C>A Intron 1	5620 A>C 3' UTR
1		
7		
24		
46		

Oromo

Haplotype id	2159 G>A Intron 4	5347 T>C Exon 7 (N516N)
5		
7		
16		
61		

2.3.5.4 *CYP1A2* cds haplotypes

In order to restrict the haplotype set to those most likely to affect the structure/function of the enzyme, haplotypes were constructed using only non-synonymous polymorphisms from samples with full genotypes. A total of 13 *CYP1A2* cds haplotypes were determined/inferred in the Ethiopian ascertainment and NIEHS populations, none of which had recombined (table 2.11). Haplotype cds 8, characterised by no mutations, was the haplotype reported in the chimpanzee. Four haplotypes (cds 1, 2, 4 and 13) were compound haplotypes with two non-synonymous alterations, except cds 1 which had three. Only 1513C>A in exon 3 (S298R) was observed in multiple haplotypes, one of which contained the premature stop codon (Y495Ter) in exon 7.

2.3.5.5 Predicted effect of amino acid substitutions on CYP1A2 structure/function

The possible impact of each of the amino acid substitutions on the structure and function of CYP1A2 was performed using PolyPhen software. The predicted effect of each amino acid change is shown in table 2.11 whilst the predicted effect of each cds haplotype, based upon the single amino acid alterations, is shown in table 2.12 (310 G>A in exon 2 (D104N) and 1460 C>T in exon 3 (R281W) could not be incorporated into haplotypes because they were not polymorphic in samples without missing genotypes). Haplotypes cds 3, 6, 7, and 9 - 12 were predicted to have no effect on the structure/function of the protein, and consequently may not differ from the ancestral haplotype (cds 8). Haplotype cds 9, harbouring T438I and previously reported by the CYP450 Allele Nomenclature Committee as *CYP1A2**14, has been shown to have a *CYP1A2**1A-like function (Murayama et al., 2004). No functional studies have however been reported concerning the effects of the other amino acid changes. Haplotypes cds 1, 2, 4, 5 and 13 were predicted to be damaging in some way.

Table 2.11 The predicted effect of the non-synonymous *CYP1A2* variants on the structure and function of the protein using PolyPhen software

CYP1A2 variant	Location	Amino acid change	Amino acids	Predicted effect on CYP1A2 structure/function
53C>G	Exon 2	S18C	Serine > Cysteine	Possibly damaging
217G>A	Exon 2	G73R	Glycine > Arginine	Probably damaging
310G>A	Exon 2	D104N	Aspartic Acid > Asparagine	Benign
331C>T	Exon 2	L111F	Leucine > Phenylalanine	Benign
613T>G	Exon 2	F205V	Phenylalanine > Valine	Probably damaging
1460C>T	Exon 3	R281W	Arginine > Tryptophan	Probably damaging
1513C>A	Exon 3	S298R	Serine > Arginine	Benign
3463C>T	Exon 6	T395M	Threonine > Methionine	Benign
3468A>C	Exon 6	N397H	Asparagine > Histidine	Benign
5094T>C	Exon 7	F432S	Phenylalanine > Serine	Probably damaging
5105G>A	Exon 7	D436N	Aspartic Acid > Asparagine	Benign
5112C>T	Exon 7	T438I	Threonine > Isoleucine	Benign
5253C>G	Exon 7	P485R	Proline > Arginine	Possibly damaging
5284C>A	Exon 7	Y495Ter	Tyrosine > STOP	Presume definite damage
5328G>A	Exon 7	R510Q	Arginine > Glutamine	Benign

Table 2.12 Haplotype inference across the cds exons (only non-synonymous changes) in the Ethiopian ascertainment and NIEHS populations. Frequencies are shown in figure 2.7.

Nucleotide change [1]		53 C>G	217 G>A	331 C>T	613 T>G	1513 C>A	3463 C>T	3468 A>C	5094 T>C	5105 G>A	5112 C>T	5253 C>G	5284 C>A	5328 G>A	Predicted effect on CYP1A2 structure/function [3]
Location		Exon 2	Exon 2	Exon 2	Exon 2	Exon 3	Exon 6	Exon 6	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	
Amino acid change		S18C	G73R	L111F	F205V	S298R	T395M	N397H	F432S	D436N	T438I	P485R	Y495Ter	R510Q	
Haplotype id [2]	cds 1														Probably damaging
	cds 2														Presume definite damage
	cds 3														Undamaged
	cds 4														Possibly damaging
	cds 5														Probably damaging
	cds 6														Undamaged
	cds 7														Undamaged
	cds 8														*CYP1A2**1A - like function
	cds 9 (*14)														Undamaged
	cds 10														Undamaged
	cds 11														Undamaged
	cds 12														Undamaged
	cds 13														Possibly damaging

[1] Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1) from the CYP1A2 genomic reference sequence (NC_000015.8)

[2] White cell, allele observed in CYP1A2*1A, grey cell, derived allele. Ten haplotypes (shown in bold) were unambiguously resolved from homozygous genotypes at all loci or from a single site heterozygote

[3] Predictions made using PolyPhen software. Predicted effects of each cds haplotype are based upon the single amino acid alterations

Haplotype cds 9 was reported in *CYP1A2**14 (*CYP1A2**1A-like function) by the Human CYP450 Nomenclature Committee. 5112 C>T (T438I) was observed in an East Asian sample but was excluded from *CYP1A2* (entire gene) analysis due to missing data at other SNP loci in that individual.

2.3.5.6 *CYP1A2* cds haplotype frequencies

CYP1A2 cds haplotype frequencies are shown in figure 2.7. The modal haplotype (≥ 83 %) in all populations was cds 8 (ancestral). In fact, this was the only haplotype in Europeans and Hispanics. Notably, potentially damaging cds haplotypes were only observed in Amhara, Anuak, Oromo and Yoruba, and their frequencies never exceeded 8 % in any one group. All chromosomes found more than once in any one NIEHS population, were detected in the Ethiopian ascertainment population.

2.3.5.7 *CYP1A2* cds diplotype frequencies

Diplotype configurations of *CYP1A2* cds haplotypes are shown in figure 2.8. All diplotypes included at least one cds haplotype which was predicted to have an undamaged protein and the majority had a copy of cds 8 (ancestral). The modal diplotype (≥ 67 %) was 8/8 (ancestral) in all populations. This was the only diplotype in Europeans and Hispanics.

2.3.6 *CYP1A2* diversity in the Ethiopian ascertainment and NIEHS populations

Gene diversity for the entire gene (figure 2.9) was highest in African Americans (0.91 ± 0.06) and lowest in Europeans (0.49 ± 0.08). African Americans, Yoruba, Maale and Oromo were all more heterozygous than the world dataset.

Gene diversity for the cds region (figure 2.9) was lower than that for the entire gene. Europeans and Hispanics (0.00) were the least diverse whilst Yoruba (0.32 ± 0.16) were the most diverse. Again, Yoruba, Maale and African Americans were all more heterozygous than the world dataset, but Anuak and the combined Ethiopian dataset were now more heterozygous than the world combined.

Nucleotide diversity for the entire gene (figure 2.10) was highest in Oromo (0.05 ± 0.03) and lowest in Europeans (0.03 ± 0.02). Oromo, Maale, African Americans and the combined Ethiopian ascertainment population were all more diverse than the world dataset.

Consistent with gene diversity, nucleotide diversity for the cds region (figure 2.10) was lower than that for the entire gene, and was lowest in Europeans and Hispanics (0.00) and highest in Yoruba (0.04 ± 0.0398). Yoruba, Anuak, Maale and the combined Ethiopian ascertainment population were all more heterozygous than the world dataset.

Notably however, both gene and nucleotide diversity values for NIEHS populations should be treated with caution given their small sample sizes and large variances.

Figure 2.7 *CYP1A2* cds haplotype frequencies in the Ethiopian ascertainment and NIEHS populations. Asians are East Asians.

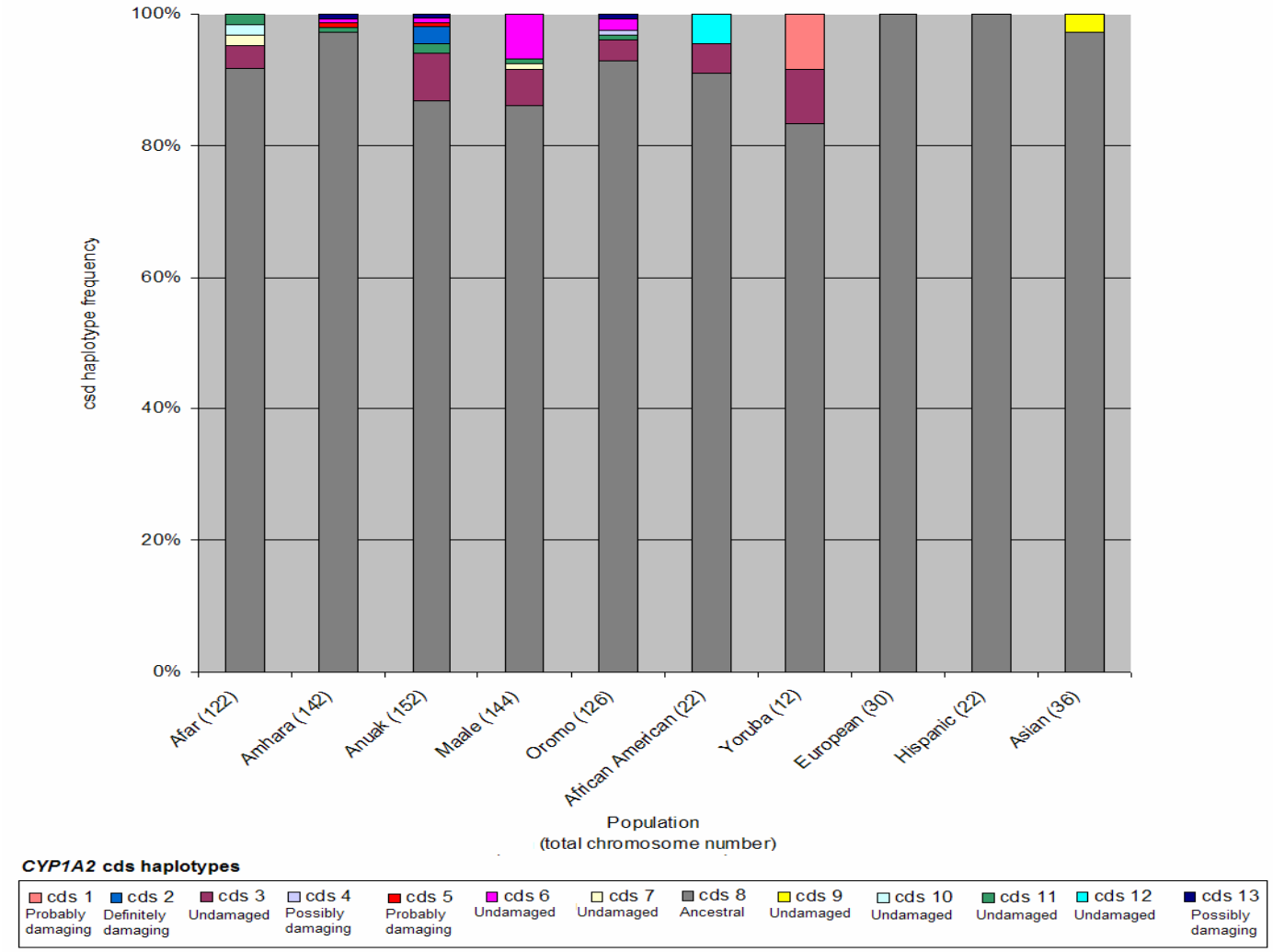


Figure 2.8 ***CYP1A2*** **cds diplotype frequencies in the Ethiopian ascertainment and NIEHS populations.** Diplotype numbers correspond to those previously assigned to each cds haplotype. Asians are East Asians.

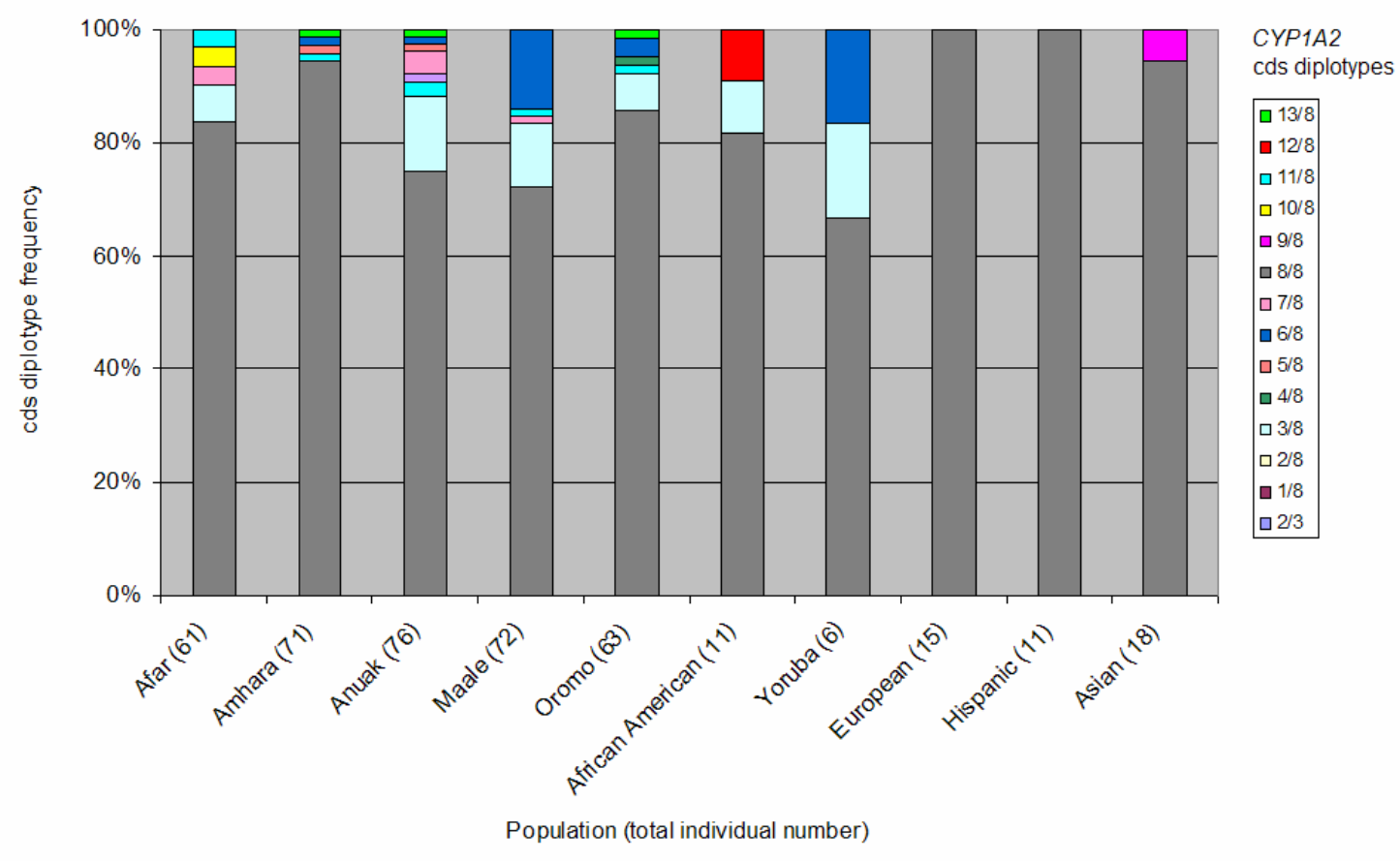


Figure 2.9 Gene diversity (h) based on the *CYP1A2* entire gene (above) and cds region (below). Error bars indicate standard deviation.

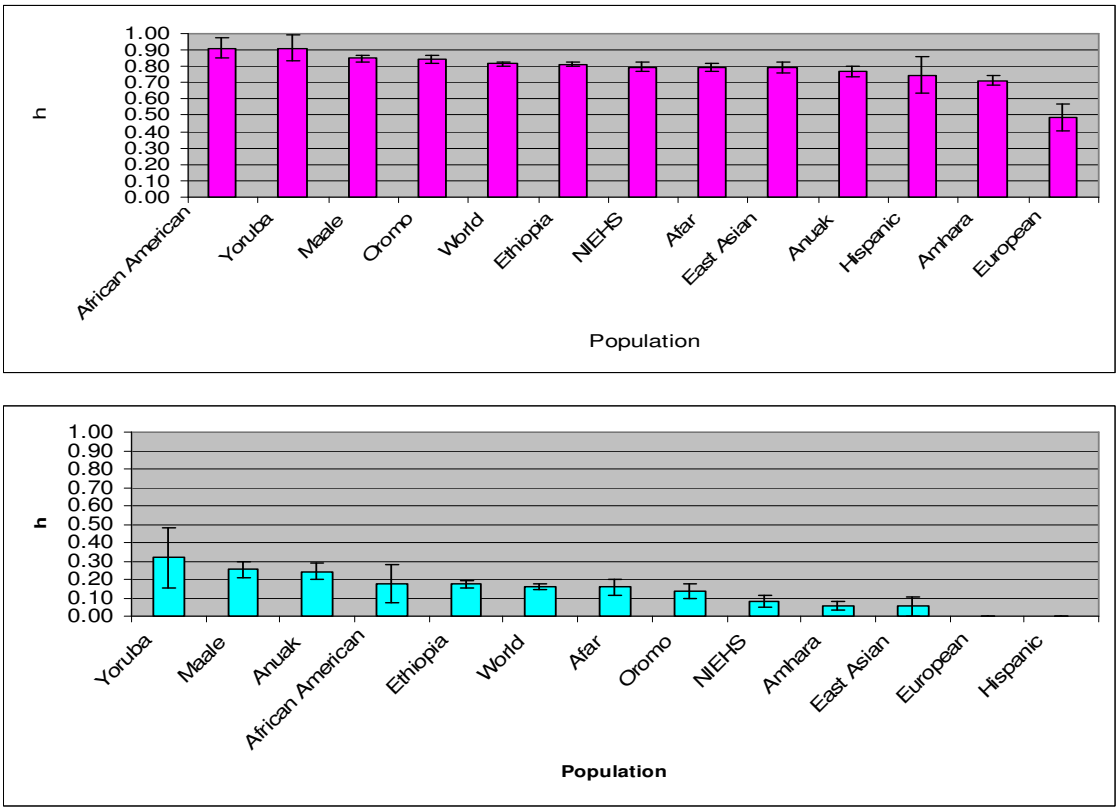
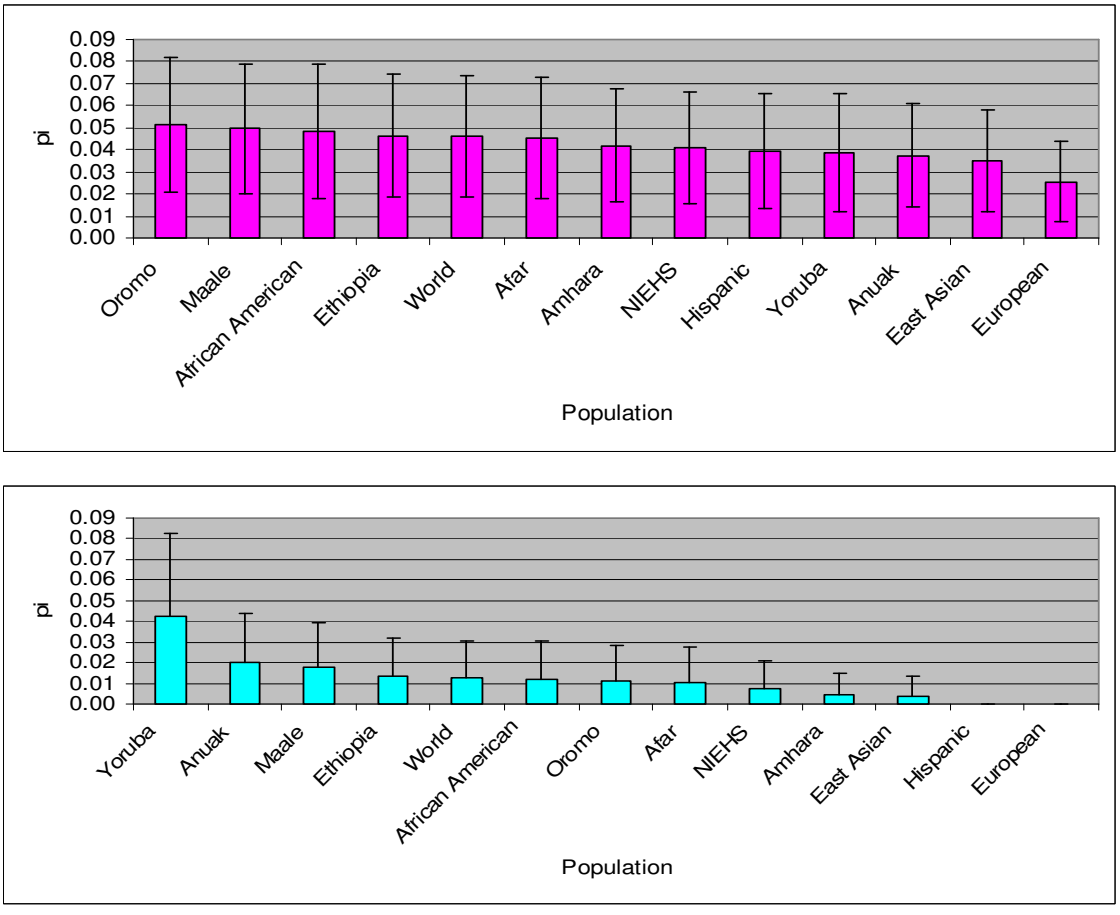


Figure 2.10 Nucleotide diversity (π or pi) based on the *CYP1A2* entire gene (above) and cds region (below). Error bars indicate standard deviation.



2.3.7 How similar are the Ethiopian ascertainment and NIEHS populations in terms of their *CYP1A2* haplotype frequencies?

The majority of populations were significantly different ($p < 0.05$) when entire gene haplotypes were considered (figure 2.11a). When the haplotype set was restricted to markers which are most likely to affect the structure/function of the protein (i.e. the cds region), considerably less pairwise differentiation was observed (figure 2.11b). Notably, significant differences were only observed amongst Ethiopian populations when the cds region was analysed (figure 2.11b).

Figure 2.11 Exact test of population differentiation p values (lower triangle) and significant/not significant (+/-) differences at the 5 % threshold (upper triangle) for *CYP1A2* entire gene (a) and cds (b) haplotypes

	(a) *CYP1A2* (entire gene) haplotypes									
	Afar	Amhara	Anuak	Maale	Oromo	African American	Yoruba	European	Hispanic	East Asian
Afar		-	+	+	-	+	+	-	-	+
Amhara	0.11		+	+	+	+	+	-	+	+
Anuak	< 0.01	< 0.01		+	+	+	-	+	+	+
Maale	< 0.01	< 0.01	< 0.01		+	+	+	+	+	+
Oromo	0.13	0.01	< 0.01	< 0.01		+	+	-	-	-
African American	< 0.01	< 0.01	0.02	< 0.01	0.02		-	+	-	-
Yoruba	< 0.01	< 0.01	0.18	< 0.01	0.05	0.96		+	+	-
European	0.48	0.61	< 0.01	0.01	0.46	< 0.01	< 0.01		-	+
Hispanic	0.16	0.03	< 0.01	0.01	0.46	0.12	0.04	0.06		-
East Asian	0.01	< 0.01	< 0.01	< 0.01	0.07	0.06	0.09	< 0.01	0.1	

	(b) *CYP1A2* cds haplotypes (non-synonymous variants)									
	Afar	Amhara	Anuak	Maale	Oromo	African American	Yoruba	European	Hispanic	East Asian
Afar		+	-	+	-	-	-	-	-	-
Amhara	0.01		+	+	-	-	-	-	-	-
Anuak	0.07	< 0.01		+	-	-	-	-	-	-
Maale	< 0.01	< 0.01	< 0.01		-	-	-	-	-	-
Oromo	0.33	0.19	0.23	0.09		-	-	-	-	-
African American	0.4	0.21	0.54	0.26	0.35		-	-	-	-
Yoruba	0.16	0.07	0.35	0.18	0.19	0.64		-	-	-
European	0.89	1.00	0.66	0.44	0.86	0.18	0.07		-	-
Hispanic	1.00	1.00	0.76	0.53	1.00	0.48	0.11	1.00		-
East Asian	0.39	0.67	0.29	0.13	0.56	0.31	0.16	1.00	1.00	

To establish the pattern of *CYP1A2* genetic structuring within and among populations in this study, hierarchical Fsts were calculated for various datasets (table 2.13). The majority of variation occurred within populations, but notably, of all the datasets, most variation (1.7 %) was observed among Ethiopians when cds haplotypes were analysed.

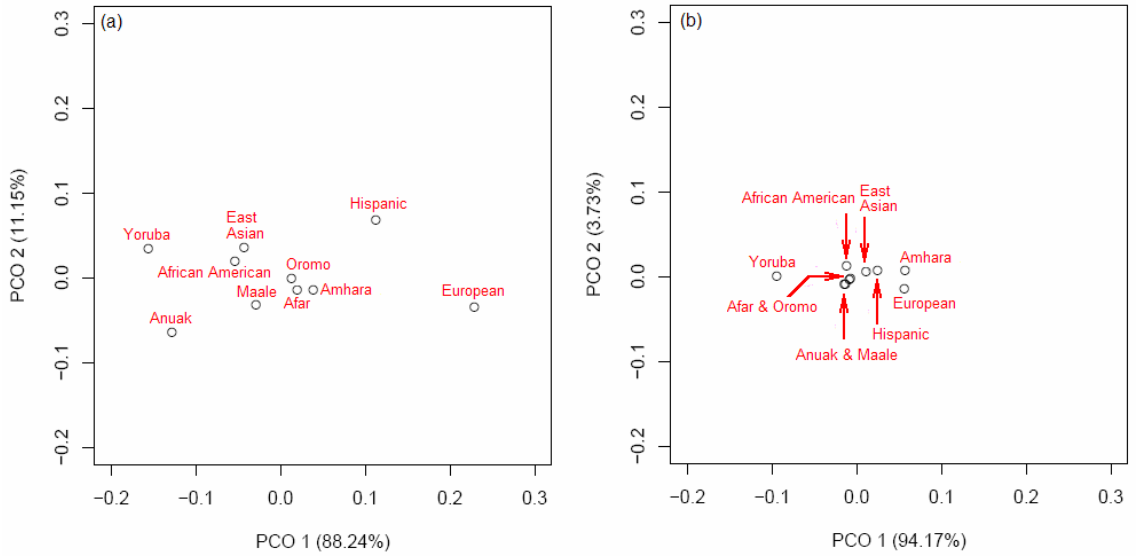
Table 2.13 Hierarchical Fsts based on *CYP1A2* entire gene (black) and cds (red) haplotypes

Dataset	FST	Variation among populations (%)	Variation within populations (%)	p value
Individual Ethiopian ascertainment & NIEHS populations	0.04 0.02	4.5 1.6	95.5 98.4	$p < 0.00001$ $p < 0.00001$
Combined Ethiopian ascertainment population & individual NIEHS populations	0.07 0.005	6.7 0.5	93.3 99.5	$p < 0.00001$ $p > 0.1$
Ethiopian ascertainment populations	0.03 0.02	3.1 1.7	96.9 98.3	$p < 0.00001$ $p < 0.00001$

PCO plots of genetic distances are shown in figure 2.12. Concerning *CYP1A2* (entire gene) haplotypes (figure 2.12a), Yoruba and Europeans were most dissimilar, with Anuak being closest to Yoruba and Hispanics being closest to Europeans. Consistent with their assumed similar mixed ancestry, Afar, Amhara and Oromo formed a tight cluster which was almost equidistant between Yoruba and Europeans. Maale were observed between the Afar, Amhara and Oromo cluster and Anuak, whilst African Americans and East Asians clustered together between the Afar, Amhara and Oromo cluster and Yoruba. The closeness of East Asians to African Americans was surprising and required further investigation with more samples. When another East Asian cohort (Japanese) (Soyama et al., 2005) was included in the analysis, African Americans were not tightly clustered with this group (data not shown).

When cds haplotypes were analysed (figure 2.12b), Yoruba were considerably differentiated from all other populations. This observation should however be treated with caution since the NIEHS Yoruba population included only 12 subjects. Similar to *CYP1A2* (entire gene) haplotype data, Yoruba and Europeans were most dissimilar while Afar and Oromo, and East Asians and African Americans were respectively close to each other. In stark contrast however, Amhara were no longer tightly clustered with Afar and Oromo but were close to Europeans, and Anuak and Maale were firmly clustered together.

Figure 2.12 PCO plots of genetic distance (Fst) among the Ethiopian ascertainment and NIEHS populations for *CYP1A2* entire gene (a) and cds (b) haplotypes. Pairwise Fsts are shown in supplementary figure S1.



2.3.8 Imputation of missing genotype data

A total of 26 % of individuals had at least one of the polymorphic nucleotides not characterised due to failed sequencing. Missing genotypes were imputed using fastPHASE software. Samples with missing data were then included in the analysis to determine whether inclusion of

imputed data would significantly affect analysis at the population level. Results were largely consistent with those from data without missing genotypes (data not shown).

2.4 Discussion

All seven exons and flanking regions of the adjacent introns of *CYP1A2* were sequenced in five Ethiopian populations: Afar, Amhara, Anuak, Maale and Oromo. This is the first detailed report of human *CYP1A2* genetic variation in Ethiopian populations. Unlike previous studies, which have only sequenced twelve Ethiopians of a mixed ethnic origin (Aklillu et al., 2003) or genotyped six SNPs in just six Ethiopians with unknown ethnicities (Jiang et al., 2006), this study resequenced the entire *CYP1A2* gene in five different Ethiopian ethnic groups. Each sample set comprised at least 76 individuals. Corresponding sequence data from an additional five populations, (African Americans, Yoruba, Europeans, Hispanics and East Asians), generated by the NIEHS SNPs programme was also included in the analysis. Sequence data was available from a Japanese population (n = 250) (Soyama et al., 2005) but was not used throughout the whole of the analysis because subjects were either patients with arrhythmia or patients with epilepsy.

The combined NIEHS sample amounted to only 95 individuals with just 15, 12, 22, 22 and 24 subjects included in African American, Yoruba, European, Hispanic and East Asian sample sets, respectively. Given the following equation: $(1 - p)^n$ = significance threshold (where n = the number of chromosomes and p = the proportion of the minor allele), the minimum number of chromosomes required to identify the proportion of the minor allele with 95 % confidence when the minor allele frequency is 1 %, 2 %, 3 %, 4 % and 5 % is 298, 148, 98, 73 and 58 respectively. Hence, the minimum number of individuals required to identify alleles at frequencies of 1 %, 2 %, 3 %, 4 % and 5 % with 95 % confidence is 149, 74, 49, 36 and 29 respectively. The Ethiopian ascertainment sample sets are consequently large enough to identify variants at frequencies of 2 % or above in each of the populations with 95 % confidence. However, the largest sample set used in the NIEHS study was 24 individuals (East Asians). Consequently, none of the NIEHS sample sets are large enough to detect variants present at 5 % in the population with 95 % confidence. Small sample sizes such as these, which have been used in the majority of *CYP1A2* sequencing studies reported to date, lack power and prevent reliable conclusions being reached. In spite of this however, the NIEHS data proved to be useful in placing the Ethiopian data, albeit tentatively, into a worldwide context. Future studies designed to analyse genetic sequences at the population level should include a minimum of 74 people, or 148 chromosomes, in the sample population for a chance of capturing polymorphisms ≥ 2 % with 95 % confidence. This number also allows the data to be analysed in terms of both chromosomes and people which increases its usefulness.

The main finding of this study is that it is clear that there was a substantial amount of previously unreported *CYP1A2* genetic variation present in the Ethiopian sample sets. A total of 30 novel

CYP1A2 mutations were identified in the Ethiopian ascertainment populations alone. Haplotype analysis, where missing data was not imputed, revealed no less than 64 haplotypes covering the entire *CYP1A2* gene in the combined Ethiopian ascertainment and NIEHS populations. Only three of these have been reported by the Human CYP450 Nomenclature Committee as *CYP1A2**1B, *1F and *1M meaning that a total of 61 novel haplotypes, many of them rare, were identified in this study. Notably, 48 of these novel haplotypes were specific to Ethiopia.

When haplotypes were constructed using only non-synonymous polymorphisms so as to restrict the haplotype set to those most likely to affect the protein (although it is accepted that variation in splice sites could also affect the protein's structure and variation in the promoter could affect gene expression), 13 haplotypes were identified. Only two of these had been reported by the Human CYP450 Nomenclature Committee (the ancestral haplotype and *CYP1A2**14 with T438I). Consequently eleven haplotypes with novel amino acid changes (eight specific to Ethiopia) were identified by this study.

When missing data was imputed, 85 haplotypes (covering the entire *CYP1A2* gene) were identified. With the exception of *CYP1A2**1B, *1F and *1M, none of these haplotypes have previously been reported. Again the majority (61) were only identified in Ethiopians, all of which were novel. Analysing only non-synonymous variants, two additional haplotypes, compared with the unimputed data analysis, were identified. These were identified in the NIEHS European population but have not been reported by the Human CYP450 Nomenclature Committee. It is thus evident that not only has variation revealed by resequencing in Ethiopian populations yielded additional information, but that not all variation recorded in publicly available databases has yet been reported to the Human CYP450 Nomenclature Committee. Prior to this study most resequencing of the entire *CYP1A2* gene has been of Japanese samples. Of all the haplotypes with novel amino acid changes identified in this study, only one has been previously documented by the Human CYP450 Nomenclature Committee (cds haplotype 9 (*CYP1A2**14) identified only in the NIEHS East Asian population). It is clear that *CYP1A2* Allele Nomenclature must be updated regularly if it is to fully deliver its potential utility for pharmacogenetic prediction.

Consistent with the hypothesis that there is more human genetic variation in sub-Saharan Africa than in the rest of the world combined, *CYP1A2* gene diversity and nucleotide diversity were always observed to be highest in African populations in this study. Furthermore, Maale (1994 census population 46,458), Oromo (1994 census population, 17,080,318), Anuak (1994 census population 45,665), and the combined Ethiopian ascertainment population were often more diverse than the world dataset. Gene diversity and nucleotide diversity values were also comparatively high in both the NIEHS African American and Yoruba data sets.

Exact tests of population differentiation revealed considerably fewer differences among *CYP1A2* cds haplotypes than among *CYP1A2* entire gene haplotypes. This finding supports evolutionary conservation of the *CYP1A2* coding region and emphasizes the importance of the

gene in nature (refer to chapter 3 for analyses regarding selection). Consistent with other studies which show Ethiopia's marked variability relative to other global populations, the Ethiopian ascertainment populations were the only ethnic groups to be differentiated when cds haplotypes (only non-synonymous changes) were considered. In light of this, the general Ethiopian population should perhaps not be treated, at the CYP1A2 protein level, as one homogenous group, a finding which undoubtedly has implications for future therapeutic intervention in Ethiopia. In this study, about 3 % of *CYP1A2* genetic variation was evident among populations. This finding is also of therapeutic importance because statistically significant variation exists among indigenous groups living in close geographical proximity.

Consistent with anatomically modern humans migrating out of Africa via Ethiopia, and a more recent migration of Semitic speaking peoples from Arabia into Ethiopia, much of the *CYP1A2* variation found outside Ethiopia remains present within Ethiopian groups. Furthermore, all of the common variation observed in the NIEHS populations (in the regions sequenced in the Ethiopians in this study) was detected in the Ethiopian ascertainment population. Consequently, the Ethiopians could perhaps serve not only as a suitable population for the development of *CYP1A2* diagnostic markers/tests useful in pharmacogenetic prediction in populations worldwide but also to ensure that such tests were suitable not only for developed countries. These findings also highlight the need to conduct population genetic research in Ethiopians if conclusions reached concerning populations outside of Ethiopia are to be interpreted in context.

Several of the *CYP1A2* alleles identified in this study were predicted to change the structure/function of the protein. At the cds level (only non-synonymous variants), six novel haplotypes were predicted to alter the functional activity of the protein. At the entire *CYP1A2* gene level, novel haplotypes 49 and 66, which stem from *CYP1A2**1K, are predicted to have a reduced function. Since all variants predicted to change the function of the protein were observed in individuals who were at least 18 years old (the age of the Yoruba individual (DY10) with cds haplotype 1 (predicted to be damaging) is not reported but the subject has a biological child), it is clear that these variants are compatible with maintenance of life to reproductive age. Tolerance of functional variation within *CYP1A2* is consequently evident. Haplotype cds 2 (with S298R and the premature stop codon Y495Ter) was identified in Anuak at 3 % (with a 95 % confidence interval of 0.007 to 0.066 (exact Pearson-Klopper method)), hence in a population numbering 45, 655 (the 1994 census record for Anuak), it is expected that 2657 people would carry one copy of the premature stop codon whilst 41 people would carry two copies and potentially have a non-functional CYP1A2 protein (providing such a condition is not homozygous lethal). In light of this, it is evident that future studies are warranted regarding the functional significance of the premature stop codon in exon 7.

All of the other haplotypes that were predicted to have a damaging effect on the protein were very rare and never exceeded more than a single observation in any one ethnic group. It is obvious that unrecognised variation cannot be studied *in vivo*, and paucity of such knowledge

may lead to inappropriate therapeutic intervention and increase the risk of adverse drug reactions. However, should public health workers be concerned about how the non-modal ~ 3 % in a population react to a drug given that at least 67 % of each population in this study did not carry a non-synonymous mutation? Similarly is it appropriate to utilise resources to investigate the impact of haplotype 61 (*CYP1A2**1F), which may have a higher inducibility compared to *CYP1A2**1A, given its rarity in Oromo and Hispanics? Answers to these questions fall within the realm of pharmacoeconomics and pharmacoethics and are therefore outside the confines of this study.

Of considerable practical relevance in healthcare, only these haplotypes: haplotype 1 (*CYP1A2**1B), haplotype 5 (*CYP1A2**1M) and haplotype 7 were modal in any population and together never fell below 46 %. Of even more relevance, the three haplotypes did not vary so far as non-synonymous mutations are concerned and consequently are not likely to differ from the *CYP1A2**1A in terms of protein function. This finding is important for public health policy formers since most people in all populations in this study may be expected to have normal CYP1A2 function. Moreover, all cds diplotypes included at least one cds haplotype which was predicted to code for a normal protein and the vast majority of people carried at least one copy of the unmutated haplotype (cds 8). Nevertheless, given the frequency of the non 'wild type' cds haplotypes there will be individuals, in different proportions in different ethnic groups, that may be expected to have two copies of non 'wild type' cds haplotypes. But in this study in no case is this predicted to be > 2 %. In view of this, perhaps a large proportion of the populations have a normal, or 'wild type'-like, CYP1A2 function overall, even if individuals do carry one damaged copy of the protein.

In terms of non-synonymous variants, Europeans, Hispanics and East Asians were considerably less variable than Ethiopians, African Americans and Yoruba. As a consequence, public health policy makers may not have to be concerned about variable drug response, due to variation in the protein, in non-African populations. However given that currently most drug testing is undertaken on non-African populations and that Yoruba and Europeans were always most dissimilar in terms of *CYP1A2* haplotypes, more testing on non European/Asian populations is warranted. With increasing numbers of people having a recent African descent living in Europe and the Americas their pharmacogenetic profiles should be represented in clinical trials. In addition, there should be close attention paid to them in post-marketing surveillance and greater awareness of variability amongst them.

Studies of the range of *CYP1A2* variants observed in the Ethiopian ascertainment populations will undoubtedly prove useful across different scientific disciplines. The vast majority of mutations were found in the non-coding regions of the gene. Although it is important to consider functional markers, particularly for healthcare utility, non-functional markers are important in population studies including association studies investigating pharmacokinetic variation due to the cis promoter region, not least in *CYP1A2* where it has not yet been fully characterised. Similarly, non-functional markers should prove useful in investigating copy

number variation and variability due to mutations in non-coding regions e.g. splice sites, initiation sites.

From this study, it is evident that varying the choice of markers can reveal different levels of inter-population differentiation and intra-population structure. For example, when the cds haplotypes were considered, in stark contrast to the entire gene haplotypes, Amhara were no longer tightly clustered with Afar and Oromo but were close to Europeans. Similarly, Anuak and Maale were placed close together. Variants confined to Ethiopia may be particularly useful in the study of migration and ethnic group relationships in Ethiopia. Among the non-synonymous variants 5284C>A in exon 7 is particularly interesting since this is the first report of a premature stop codon in this gene. Apart from the splice site variant in intron 6, identified as a singleton in Caucasian (Allorge et al., 2003) and Japanese subjects (http://www.pharmgkb.org/), there have been no other reports of mutations altering the primary structure of the protein and leading to a truncated non-functional enzyme. This observation prompts the question of whether mutations causing the protein to be truncated are simply not recognised at appreciable levels in the population, because if they do occur in a homozygous form, foetuses do not survive. The data generated in this study is now being analysed to assess the probability that truncated proteins due to stop codons, or altered splice sites in introns 1-6 are not observed because their presence would be lethal *in utero*. More samples should also be sequenced to try and establish whether the absence of mutations in splice sites in this study was a consequence of small sample size. Functional studies should also be able to determine whether the premature stop codon in exon 7 does actually lead to a non-functional enzyme or a protein with reduced function. If non functionality is the case and if homozygotes do exist, then such individuals would be living human *CYP1A2* knock outs whose existence would open interesting possibilities for research into P450 mediated metabolic activity.

In this study, as in others, 1879C>A in exon 3, resulting in S298R, and the deletion in the 3' UTR (4714Gdel) were only observed in African populations or populations with recent African ancestry. With the exception of 1879C>A observed as a singleton in a Hispanic sample (http://snp500cancer.nci.nih.gov/), there have been no reports of these mutations in non-Africans. Both SNPs may therefore be of practical use in anthropological research.

Several recombination events were inferred in *CYP1A2*, particularly towards the 3' end of the gene (from intron 6 onwards). Considering this, and given that the premature stop codon, splice site variant and several potentially damaging amino acid substitutions have also been found in exon 7, perhaps this region, particularly the 3' UTR, is not as functionally important as is the rest of the gene. In turn, this finding questions whether two of the enzyme's catalytic residues are in fact located in exon 7, as a recent study suggests (Sansen et al., 2007). Consistent with this possibility, prior to this study variation has not been reported in exon 1 and very few non-synonymous variants have been identified in exons 4 and 5. This is not surprising given that exon 1 contains the leader sequence of the mRNA, thus playing a crucial role in protein expression, and exon 4 is thought to contain part of the enzyme's active site (Sansen et

al., 2007). Variation was also not observed in exons 1, 4 and 5 in this study, further suggesting that the beginning and middle of the gene may be more functionally significant than the end. However, is what appears to be recombination, in fact a result of gene conversion? Distinguishing between the two phenomena is not currently possible with the data generated in this study. The plethora of variation identified in this study should prove of considerable utility in designing association studies to elucidate the effects of non-coding sequence variation on the activity of CYP1A2.

2.5 Conclusion

Drugs thought to be metabolised by CYP1A2 are widely administered to Ethiopians. However in comparison with other countries, particularly Japan, studies investigating the distribution of human *CYP1A2* genetic variation have been scarce. Those that have been undertaken have been limited to few populations and small sample sizes. This study has contributed to correcting this imbalance and found that Ethiopia has more *CYP1A2* variation than other populations characterised to date and, in some respects, the rest of the world combined. It also evidences much of the variation found on a global scale. Not only does this serve as further support for the proposition that anatomically modern humans migrated out of Africa from Ethiopia, but also emphasizes the value of conducting population genetic research with Ethiopians if appropriate conclusions are to be formulated concerning populations outside of Ethiopia. This study found a substantial amount of uncharacterised variation in Ethiopia. Unrecognised variation can lead to unsuitable therapeutic, prophylactic and diagnostic intervention and can increase the risk of an adverse drug reaction. Investigations such as this are not only of benefit to the indigenous populations of Ethiopia, but are also of increasing importance in directing public healthcare policies in the developed world, where the number of individuals of recent Ethiopian descent is growing.

3 The recent evolutionary history of *CYP1A2*

3.1 Introduction

Diversity in the human *CYP1A2* gene, which was described in chapter 2, could have been shaped by several different evolutionary forces. These processes include mutation and recombination which generate new genetic variation, selection which shapes pre-existing variation and genetic drift which serves to remove variation. The interplay of these different forces, within the framework of the demographic histories of various population groups, in the evolutionary history of *CYP* genes is likely to have been very complex. *CYP1A2* detoxifies environmental xenobiotics in mammals and birds and is thought to have evolved 350 million years ago as the result of a duplication of *CYP1A1* (Heilmann et al., 1988; Liang et al., 1996). Exons 2, 4, 6 and especially 5 are strikingly conserved between *CYP1A2* and *CYP1A1* in both nucleotide types and total number of bases (Ikeya et al., 1989). Little is known however about the preservation of *CYP1A2* among human populations. This chapter attempts to redress this imbalance and focuses on extracting information about the evolutionary history of *CYP1A2* from the genetic variation observed in the Ethiopian ascertainment and NIEHS populations.

A substantial amount of *CYP1A2* genetic variation was observed in the Ethiopian ascertainment populations (see chapter 2). The depth of this unique data set allows a detailed investigation into the evolution of *CYP1A2* in humans. Network analysis of *CYP1A2* haplotypes will provide insight into *CYP1A2* evolution in human populations. Furthermore, comparisons between *CYP1A2* sequences in humans and some of their closest living relatives (other ape species) will root the network, thus providing further insights into the evolutionary history of the gene. A rooted network can be used to construct a relative chronology for genetic changes within *CYP1A2*. For example, changes within the external branches of the network must have occurred after those found closer to the root within the same clade. A rough absolute chronology can also be produced by dating of the *CYP1A2* variants. Such contextualisation of *CYP1A2* genetic data may prove useful in the reconstruction of human pre-history. Time estimates can place genetic changes in a wider context and can be integrated with evidence from other disciplines attempting to examine human evolution. For instance, dates of mutations may be related to periods in human history (Thomas et al., 1998).

Positive, balancing and purifying (negative) selection can shape genetic diversity in a number of ways and whilst it can be directly detected in both intragenic and extragenic sequences, its signature can also be observed in markers linked to the locus under selection. Detecting selection is however inherently difficult and many studies lack power due to the lack of diversity observed. The Ethiopian ascertainment data set may be rich enough to detect possible signatures of selection.

3.1.1 Aims

1. Construct a rooted *CYP1A2* network for the combined cds region and, to the extent that it has a single genealogy, the entire gene.

2. Determine to what extent the combined cds region and the entire *CYP1A2* gene have been conserved among human populations and throughout primate evolution.

3. Apply tests to see if the gene is under selection, and of what type.

4. Estimate when various polymorphisms arose in *CYP1A2*.

3.2 Methods

3.2.1 *CYP1A2* sequence data

The Ethiopian ascertainment and NIEHS *CYP1A2* sequences described in chapter 2 were used in this study. Missing data was excluded from the analysis and haplotypes were inferred by the ELB approach (Excoffier et al., 2003) from the combined Ethiopian ascertainment and NIEHS population. The imputed dataset was not used in this study. Refer to the methods section in chapter 2 for details regarding sample collection, DNA extraction, sequencing and haplotype inference.

CYP1A2 sequences from the chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmeus*) and Rhesus macaque (*Macaca mulata*) were obtained from Ensembl (52: Dec 2008). *CYP1A2* in the chimpanzee was referred to as Q9N256_PANTR, whilst *CYP1A2* in the orangutan and macaque was referred to as CP1A2_PONPY and Q9N253_MACMU respectively.

3.2.2 *CYP1A2* mutation networks

Mutation networks were constructed using Network software, version 4.510 (fluxus-engineering.com). Median joining networks were constructed to limit levels of reticulation (Bandelt et al., 1999). The algorithm used to construct these median joining networks is based on the limited introduction of likely ancestral sequences/haplotypes into a minimum spanning network of the observed sequences. These likely ancestral sequences are identified through the calculation of median haplotypes. The resultant networks were drawn using Network publisher, version 1.1.0.7 (Fluxus Technology Ltd).

3.2.3 Testing for selection in *CYP1A2*

3.2.3.1 Tajima's test of neutrality *

This test calculated the *D* test statistic proposed by Tajima (1989), equation 38, for testing the null hypothesis that all mutations were selectively neutral (Kimura, 1983). The D test was based on the differences between two estimates of theta (θ), one based on the number of segregating sites (θs), and the other based on the average number of nucleotide differences between pairs for *CYP1A2* sequences (θπ). The confidence limits of *D* (two tailed test) were obtained from table 2 of Tajima (1989). These confidence limits assumed that *D* followed a beta distribution described by equation 47 in Tajima (1989).

3.2.3.2 McDonald-Kreitman test of neutrality *

This is a test of the neutral hypothesis (Kimura, 1983) proposed by McDonald and Kreitman (1991). The test was based on a comparison of synonymous and non-synonymous variation within and between species. Under neutrality, the ratio of non-synonymous to synonymous fixed substitutions (differences) between species should be the same as the ratio of non-synonymous to synonymous polymorphisms within species. The selected non-human species was the chimpanzee. P values were obtained from two-tailed Fisher's exact tests.

3.2.3.3 Fu and Li's tests of neutrality (with an outgroup) *

This test calculated the statistical tests D and F which were proposed by Fu and Li (1993) for testing the null hypothesis that all mutations were selectively neutral (Kimura 1983). Interspecific difference was obtained by comparison with the chimpanzee.

The D test statistic was based on the differences between the total number of mutations in external branches of the genealogy and the total number of mutations (equation 32 from Fu and Li (1993)). The F test statistic is based on the differences between the total number of mutations in external branches of the genealogy and the average number of nucleotide differences between pairs of *CYP1A2* sequences (page 702 of Fu and Li (1993)).

The critical values (two-tailed test) shown in tables 2 and 4 from Fu and Li (1993) were used to determine the statistical significance of the D and F test statistics.

Statistical tests of neutrality marked with * were performed on data from each population using DNAsp software (http://www.ub.edu/dnasp/).

3.2.3.4 Testing for evidence of purifying selection at radical non-synonymous SNP sites

For a given SNP site, gene diversity (heterozygosity) was estimated according to unbiased formulae of Nei (1987) (see methods in chapter 2).

For a given SNP site, genetic distance between each pair of populations was estimated by the formula: $d = 1-((x_1y_1)^{1/2}+(x_2y_2)^{1/2})$, where x_1 and y_1 are the frequencies of the first allele in each of the two populations respectively, and x_2 and y_2 are the frequencies of the second allele in each of the two populations respectively. The average of *d* over all loci is the average genetic distance (Nei, 1987).

SNPs were classified with respect to location and predicted effect on protein function as follows: (1) SNPs located in the 5' non-coding region outside the 5' UTR (exon 1*), (2) SNPs located in the introns, (3) SNPs located in the 3' UTR, (4) synonymous SNPs in exons, (5) non-synonymous SNPs in exons causing a conservative amino acid change, (6) non-synonymous SNPs in exons causing a radical amino acid change and (7) nonsense SNPs in exons causing premature stop codons. Radical non-synonymous changes included two SNPs (217G>A in exon 2 causing G73R, and 5094T>C in exon 7 causing F432S), each of which caused an amino acid replacement involving two amino acids with a pairwise stereochemical difference > 3.0 according to the scale described by Miyata et al. (1979) (figure 3.1). All other non-synonymous SNPs were categorised as conservative non-synonymous SNPs since they resulted in an amino acid replacement involving two amino acids with a pairwise stereochemical difference < 3.0.

* Exon 1 was monomorphic in Ethiopian ascertainment and NIEHS populations (see chapter 2).

Figure 3.1 Pairwise amino acid stereochemical differences based on amino acid residue and volume (Miyata et al., 1979)

	C	P	A	G	S	T	Q	E	N	D	H	K	R	V	L	I	M	F	Y
P	1.33																		
A	1.39	0.06																	
G	2.22	0.97	0.91																
S	1.84	0.56	0.51	0.85															
T	1.45	0.87	0.9	1.7	0.89														
Q	2.48	1.92	1.92	2.48	1.65	1.02													
E	3.26	2.48	2.46	2.78	2.06	1.83	0.84												
N	2.83	1.8	1.78	1.96	1.31	1.4	0.99	0.85											
D	3.48	2.4	2.37	2.37	1.87	2.05	1.47	0.9	0.65										
H	2.66	2.15	2.17	2.78	1.94	1.32	0.32	0.96	1.29	1.72									
K	3.27	2.94	2.96	3.54	2.71	2.1	1.06	1.14	1.84	2.05	0.79								
R	3.06	2.9	2.92	3.58	2.74	2.03	1.13	1.45	2.04	2.34	0.82	0.4							
V	0.86	1.79	1.85	2.75	2.15	1.42	2.13	2.97	2.75	3.4	2.11	0.7	2.43						
L	1.65	2.7	2.76	3.67	3.04	2.25	2.7	3.53	3.49	4.1	2.59	2.98	2.62	0.91					
I	1.63	2.62	2.69	3.6	2.95	2.14	2.57	3.39	3.37	3.93	2.45	2.84	2.49	0.85	0.14				
M	1.46	2.36	2.42	3.34	2.67	1.86	2.3	3.13	3.08	3.69	2.19	2.63	2.29	0.62	0.41	0.29			
F	2.24	3.17	3.23	4.14	3.45	2.6	2.81	3.59	3.7	4.27	2.63	2.85	2.47	1.43	0.63	0.61	0.82		
Y	2.36	3.12	3.18	4.08	3.33	2.45	2.4	3.22	3.42	3.95	2.27	2.42	2.02	1.52	0.94	0.86	0.93	0.48	
W	3.34	4.17	4.23	5.23	4.3	3.5	3.42	4.08	4.39	4.88	3.16	3.11	2.32	2.51	1.73	1.72	1.89	1.11	1.04

Significance tests (t-test, Kruskal-Wallis test and Wilcoxon matched-pairs signed-ranks test) were performed using GraphPad InStat version 3.00 for Windows 95, GraphPad Software, San Diego California USA, www.graphpad.com.

3.2.3.5 Testing for evidence of purifying selection at conservative non-synonymous SNP sites

CYP1A2 pairwise alignments between human and mouse (*Mus musculus*), and human and chimpanzee (*Pan troglodytes*) DNA sequences were obtained from Ensembl (52: Dec 2008). Since the possibility of variability at each of the conservative non-synonymous SNP sites in the mouse and chimpanzee could not be addressed given the available data, it was assumed that the mouse and chimpanzee sequences represented the more common allele in these species. In accordance with maximum parsimony (Fitch, 1971), it was also assumed that a residue representing the more common allele in human, mouse and chimpanzee has been conserved since the most recent common ancestor of the three species.

3.2.4 Genohaplotyping of an AC microsatellite and a G>C SNP (rs11072507) using a SNPstr system

The AC microsatellite was situated 5.6 kb downstream of the 3' end of CYP1A2 and rs11072507 was a G>C SNP situated 218 bases downstream of the AC microsatellite.

3.2.4.1 SNPstr assay design

A 384 bp region containing the AC microsatellite and G>C SNP (rs11072507) was amplified using the forward primer TCTCATCTCGCAACTGGGGA and the reverse primer GGGTTGGGGGCCCATTGTCS (see figure 3.2).

Figure 3.2 Primers (blue) used in the SNPstr system which incorporated an AC microsatellite (green) and the rs11072507 G>C SNP (S)

CCTTGGGCCTGTGACATAAACTTTCTGAGTCTAGACTTTCTCATCTCGCAACTGGGGATGATAATCTTAT
CTTACAGGTTGGGTGGGAGGTTTAATGAGATAATGTGTGTAAAGTGCCAAATTACAGGTCAGGGTACAAC
ACACACACACACACACACACACACACACACACACACACACACACACACCAGAAATCCTGAGAAAGGTGGACTCA
GGGCAAGATGGACCCCCGACACACGTGTGGCATGTGCAGCTGTGCACACACCAGGCCCACACGTGAGCCA
GCATTTCCCCAAACCCTGCGCTCTCTCACTGGGGTCTTGCCAGGGCCTGGGCTCAGCCCACTTCCTTCCA
CTGCTCCCCCCACCCAGAAGCTGGGGAGGGCTGGGGGTGGTAGCGTGCCAGG(S)AACAATGGGCCCCCA
ACCCCTTCCCGCCCATTGCGTGAGAAGGAGCTGCCGGAAGGAAACTCACCTCCCAGTGGGAC

Since the 3' end of the reverse primer annealed to the site of the G>C SNP, each allele was independently amplified. The fragment ending with the C allele was specifically amplified using

the fluorescently labelled FAM- GGGTTGGGGGCCCATTGTCG reverse primer whilst the fragment ending with the mutated G allele was specifically labelled with the HEX-GGGTTGGGGGCCCATTGTCC reverse primer. A 3' mismatch (at the penultimate base) was incorporated into both reverse primers to increase specificity of the assay. Each fluorescently labelled PCR product encompassed the SNP at one end and the microsatellite at the other and the length of the PCR product varied among chromosomes, depending solely on the number of microsatellite repeats. Consequently, the gametic phase for the SNP and microsatellite could be empirically determined by electrophoresis on a genetic analyser using fluorescent detection.

3.2.4.2 SNPstr assay protocol

Individual sample DNAs were amplified separately with each allele-specific, fluorescently labelled reverse primer and the same forward primer. Two separate PCR reactions per individual were carried out in order to increase the reliability of the results. DNA was amplified in 96 well plates in 10 µl reaction volumes containing 1 ng of template DNA, 0.3 µM of each primer (forward and reverse), 0.13 units *Taq* DNA polymerase (HT Biotech, Cambridge, UK), 9.3 nM TaqStartTM monoclonal antibody (BD Biosciences Clontech, Oxford, UK), 200 µM dNTPs and reaction buffer supplied with the *Taq* polymerase. The cycling parameters were: 4 minutes of pre-incubation at 94 ºC, followed by 35 cycles of 30 seconds at 94 ºC, 30 seconds at 56 ºC and 30 seconds at 72 ºC , with a final elongation step for 7 minutes at 72 ºC. A 2 µl aliquot of the diluted PCR product (1 in 5 dilution) was mixed with 9.89 ul of high purity (HiDi) formamide and 0.11 ul of ROX size standard (Applied Biosystems). The mixture was heated for 4 minutes at 96 ºC and immediately cooled in ice. Samples were run on an ABI 3100 genetic analyser and analysed using GeneMapper software v4.0 (Applied Biosystems (ABI), Warrington UK). Genohaplotypes (rs11072507 genotypes and AC microsatellite haplotypes) were then recorded for each sample.

3.2.4.3 Assessing the reliability of the SNPstr assay

In order to ensure that the SNPstr assay was accurately determining microsatellite lengths (by fragment mobility), a sample of rs11072507 heterozygous individuals also had their microsatellite lengths determined by sequencing. PCR conditions were the same as previously described except that the reverse primers were not labelled. Uni-directional sequencing was performed on each allele using the unlabelled reverse primers. The fragment ending with the C allele was sequenced using primer GGGTTGGGGGCCCATTGTCG whilst the fragment ending with the G allele was sequenced with primer GGGTTGGGGGCCCATTGTCC. DNA was sequenced in 96 well plates in 15 µl reaction volumes containing 6 µl of the purified PCR product, 2.4 pm of either primer, 0.75 µl of BigDye termination mix v1.1 (Applied Biosystems (ABI), Warrington UK) and 5 µl of Better Buffer (Applied Biosystems (ABI), Warrington UK). The cycling parameters were: 25 cycles of 96 ºC for 10 seconds, 55 ºC for 5 seconds and 60 ºC for

4 minutes. Sequencing reaction products were purified by mixing 80 µl of isopropanol with each product and leaving for 10 minutes at room temperature. The mixture was centrifuged for 45 minutes at 2240 x g and the resultant supernatant was discarded by inverting the PCR plate and centrifuging for 1 minute at 13 x g. 150 µl of 70 % isopropanol was added to each well and centrifuged for 15 minutes at 2240 x g. The resultant supernatant was discarded by inverting the PCR plate and centrifuging for 1 minute at 13 x g. Samples were then dried for 5 minutes at 65 ºC. Each sample was mixed with 10 µl of high purity (HiDi) formamide, heated for 4 minutes at 96 ºC and immediately cooled in ice. Samples were then run on an ABI 3100 genetic analyser and analysed using Sequencher 4.7 software (Gene Codes Corporation, USA). Microsatellite lengths were confirmed by counting the number of AC repeats observed in the sequence chromatograms.

3.2.5 Estimating the time to most recent common ancestor (TMRCA) for *CYP1A2* variants

Under the stepwise mutation model, the average square distance (ASD) in microsatellite allele size among all sampled chromosomes from the most recent common ancestral haplotype, averaged over loci, has been shown to be linearly related to µt, where µ is the mutation rate and t the coalescence time in generations (Goldstein et al., 1995; Slatkin, 1995). The AC microsatellite alleles obtained from the SNPstr assay were used to date *CYP1A2* variants in this study. Since the gametic phase for the SNP (rs11072507) and AC microsatellite was empirically determined from the SNPstr assay for each sample, the SNP (rs11072507) was used to determine to which microsatellite haplotype the allele, which was being dated, was linked. Phase was inferred for all *CYP1A2* variant alleles and rs11072507 from the pooled Ethiopian ascertainment population (excluding missing genotype data) by the ELB approach (Excoffier et al., 2003) implemented in Arlequin software, version 3.01 (Schneider et al., 2000). When both *CYP1A2* SNP alleles were on the background of both the G and C of rs11072507, recombination was inferred. Since recombination initiates a new distribution of microsatellite alleles in the evolutionary history of the gene (overlaid on the previous distribution), these variants were dated using microsatellites on the background of each of rs11072507 C and G separately and together (where possible). Of the date estimates produced from only rs11072507 G or C alleles, the older dates were assumed to indicate the coalescent date of the SNP being dated, whilst the younger was taken as the coalescent date of the recombination event. Since recombination between identical haplotypes would not affect coalescent date estimates, recombination between identical *CYP1A2* haplotypes was not accounted for in the method.

ASD and t were calculated using Ytime software, version 2.08 (Behar et al., 2003). The software includes a set of functions written for the MATLAB programming environment. The functions are designed to infer the TMRCA of a clade on a single-locus genealogy for which microsatellite haplotype information is available. The microsatellite haplotype of the MRCA is

assumed to be known. In this study, the modal haplotype was assumed to be ancestral because of its high frequency. The TMRCA (unbiased estimate plus confidence interval) was inferred under the Simple Stepwise Mutation Model (S-SMM) of microsatellite evolution. The AC microsatellite mutation rate per generation was assumed to be 0.0005 (Farrall and Weeks, 2007). Confidence intervals were obtained on the distance between the ancestral and sampled chromosomes (ignoring uncertainty in mutation rate) from simulations involving a star-genealogy model. This model was used because most non-ancestral haplotypes were rare (in some cases most were singletons) and negative Tajima D values were observed for all Ethiopian ascertainment populations (see table 3.5), indicating that the genealogy linking the *CYP1A2* chromosomes was more like the star genealogy characteristic of rapid growth than the genealogy associated with no growth. For each generation, a time period of 32 years was assumed based on data from Tremblay and Vezina (2000).

3.3 Results

3.3.1 Network analysis of *CYP1A2* haplotypes

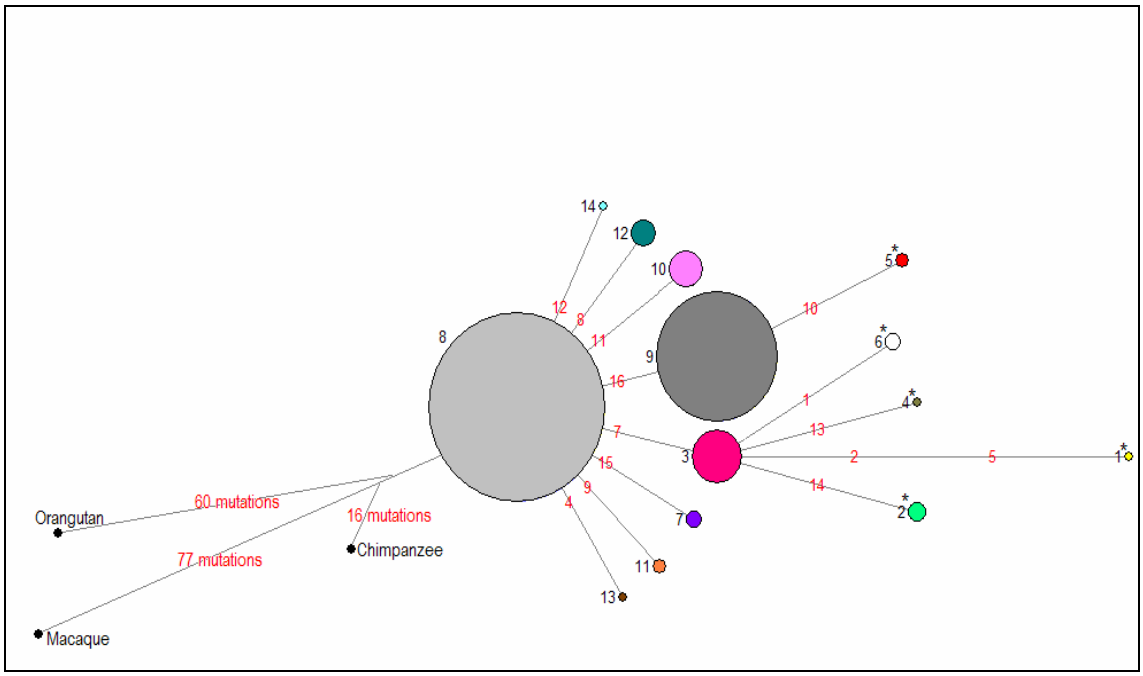
In order to establish the relationship between *CYP1A2* haplotypes observed in the Ethiopian ascertainment and NIEHS populations, mutation networks of these haplotypes were constructed. Haplotypes from the chimpanzee, orangutan and Rhesus macaque were included to determine the extent to which *CYP1A2* has been conserved throughout primate evolution. Since it was established in chapter 2 that *CYP1A2* can be divided into two LD blocks constituting the cds region and the 3' end of the gene, networks of the entire cds region (exons 2-7) were constructed in addition to networks of the entire gene. Cds haplotype networks will provide a clear insight into the conservation of the protein, whilst entire gene networks will place coding variation into the context of variation within the entire gene.

3.3.1.1 *CYP1A2* cds networks

The network of the cds region (exons 2-7) can be seen in figure 3.3. A total of 14 cds haplotypes were included in the network (see tables 3.1 and 3.2 for details regarding haplotype types and frequencies respectively). The chimpanzee was the most similar species to the human clade, followed by the orangutan and then by the macaque. The modal human haplotype (8), which was observed in all Ethiopian and NIEHS populations (table 3.2), was the most similar of the human haplotypes to the chimpanzee. The chimpanzee, macaque and orangutan haplotypes were linked to the human clade via a single common branch. Recombination was not observed in the network and all haplotypes, with one exception, within the human clade were separated by a single mutation. Most haplotypes stemmed from the modal haplotype, and as is evident from figure table 3.2, the vast majority of non-modal

haplotypes were African. Moreover, the vast majority of these were confined to Ethiopia, however this may be a reflection of the relative sizes of the datasets.

Figure 3.3 Mutation network of *CYP1A2* cds haplotypes observed in the Ethiopian ascertainment and NIEHS populations. Haplotypes from the chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus*) and Rhesus macaque (*Macaca mulatta*) are shown. Nodes represent haplotypes (shown in table 3.1) and are proportional to haplotype frequencies in the combined Ethiopian ascertainment and NIEHS population (shown in table 3.2). Mutated positions are shown in red along the grey links. All links, except those connecting the chimpanzee, orangutan and Rhesus macaque, are to scale. Haplotypes denoted with * contain amino acid alterations which are predicted to be damaging to the structure/function of the protein in some way (table 3.1).



For each of the individual exons, the chimpanzee was the most similar relative to humans, followed by the orangutan, and then the macaque (data not shown).

Exons 1, 4 and 5 were not variable in the Ethiopian ascertainment and NIEHS populations and whilst exon 1 differed from the chimpanzee by two mutations and exon 4 differed from the chimpanzee by one mutation, no variation was observed between the human and chimpanzee in exon 5. Consequently, exon 5 was the most conserved *CYP1A2* exon throughout primate evolution, followed by exon 4 and then by exon 1.

Of the exons which were variable in the Ethiopian ascertainment and NIEHS populations, exon 2 appeared to be the least conserved throughout primate evolution. A total of nine mutations separated the chimpanzee from the closest human node for exon 2, whilst only two, one and three mutations differentiated the chimpanzee from the closest human node for exons 3, 6 and 7 respectively. On the other hand however, exon 7 was the most variable of the exons and was therefore the least conserved exon among human populations.

Table 3.1 *CYP1A2* cds haplotypes used in the mutation networks shown in figure 3.3

Nucleotide change [1]		53 C>G	217 G>A	331 C>T	613 T>G	1513 C>A	3463 C>T	3468 A>C	5094 T>C	5105 G>A	5112 C>T	5253 C>G	5284 C>A	5328 G>A	5347 T>C	Predicted effect on CYP1A2 structure/function [3]
Mutated position in cds phylogenetic network		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Location		Exon 2	Exon 2	Exon 2	Exon 2	Exon 3	Exon 6	Exon 6	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	
Amino acid change		S18C	G73R	L111F	F205V	S298R	T395M	N397H	F432S	D436N	T438I	P485R	Y495Ter	R510Q	N516N	
Haplotype id [2]	1															Probably damaging
	2															Presume definite damage
	3															Undamaged
	4															Possibly damaging
	5															Probably damaging
	6															Possibly damaging
	7															Undamaged
	8															No amino acid alteration
	9															No amino acid alteration
	10															Undamaged
	11															Undamaged
	12															Undamaged
	13															Undamaged
	14															Undamaged

[1] Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1) from the CYP1A2 genomic reference sequence (NC_000015.8)

[2] White cell, non-derived allele, grey cell, derived allele

[3] Predictions made using PolyPhen software. Predicted effects of each cds haplotype are based upon the singular amino acid alterations

Table 3.2 *CYP1A2* cds haplotype frequencies in the world dataset and their distribution among the Ethiopian ascertainment and NIEHS populations

Haplotype id	Afar		Amhara		Anuak		Maale		Oromo		African American		Yoruba		European		Hispanic		East Asian		World dataset	
	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f
1	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.08	0	0.00	0	0.00	0	0.00	1	0.001
2	0	0.00	0	0.00	4	0.03	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	4	0.005
3	4	3.28	0	0.00	11	0.07	8	0.06	4	0.03	1	0.05	1	0.08	0	0.00	0	0.00	0	0.00	29	0.036
4	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
5	0	0.00	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.002
6	0	0.00	1	0.01	1	0.01	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	3	0.004
7	2	1.64	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	3	0.004
8	77	63.11	90	0.63	127	0.84	106	0.74	87	0.69	17	0.77	10	0.83	11	0.37	12	0.55	29	0.81	566	0.700
9	35	28.69	48	0.34	5	0.03	18	0.13	30	0.24	3	0.14	0	0.00	19	0.63	10	0.45	6	0.17	174	0.215
10	0	0.00	1	0.01	1	0.01	10	0.07	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	14	0.017
11	2	1.64	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.002
12	2	1.64	1	0.01	2	0.01	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	7	0.009
13	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.05	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
14	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.001
Grand Total	122	100.00	142	1.00	152	1.00	144	1.00	126	1.00	22	1.00	12	1.00	30	1.00	22	1.00	36	1.00	808	1.000

n = number of chromosomes, f = frequency

3.3.1.2 *CYP1A2* entire gene networks

The network of the whole *CYP1A2* gene is shown in figure 3.4. A total of 37 haplotypes were included in the network (see tables 3.3 and 3.4 for details regarding the types and frequencies of haplotypes respectively).

Recombination was inferred between loci 20 (2159G>A in intron 4and 34 (5347C>T in exon 7 (non-synonymous)) which manifested itself in haplotypes 16, 17, 23 and 24 (table 3.3). Recombination was also evident between loci 21 (2321G>A in intron 4) and 34, manifesting itself in haplotypes 21, 22, 23 and 24 (table 3.3). Recombined haplotypes 22 and 24, which were only observed as singletons in Afar and Oromo respectively (table 3.4), were excluded to prevent reticulations within the network shown in figure 3.4. Given their appreciable frequencies within some of the Ethiopian ascertainment and NIEHS populations (table 3.4), the other recombined haplotypes mentioned above were not excluded from the network shown in figure 3.4.

The chimpanzee, macaque and orangutan haplotypes were again linked to the human clade via a single common branch (figure 3.4). The chimpanzee was also the most similar relative to humans, but the orangutan was now the most dissimilar species to humans (figure 3.4).

Contrary to the combined cds region network, the most frequent human haplotype (33), which was observed in all Ethiopian ascertainment and NIEHS populations (table 3.4), was not the most similar human node to the chimpanzee. In this instance, haplotype 3 was the most similar and was only observed in Ethiopia in Afar, Amhara, Anuak and Oromo (table 3.4). Furthermore, the majority of haplotypes linked to haplotype 3 were only found in Ethiopian populations (table 3.4), however this may be a reflection of the relative sizes of the datasets.

Within the human clade, single mutations separated all haplotypes, with three exceptions (figure 3.4). Similar to the cds region network, the majority of entire gene haplotypes stemmed from the modal haplotype and the vast majority of non-modal haplotypes were only observed in samples with a recent African ancestry, the majority of which were confined to Ethiopia (table 3.4). Again however, this may be a reflection of the relative sizes of the datasets.

Cds haplotype 8 was observed on the background of most of the entire gene haplotypes, including the modal haplotype and haplotype 3 which was most similar to the chimpanzee (figure 3.4). Cds haplotype 9 (characterised by no mutations in table 3.1) was found on the background of the recombined entire gene haplotypes 22 and 24 (figure 3.4). Notably, most of the cds haplotype backgrounds which were predicted to be damaging in some way (see table 3.1 for phenotype predictions) were directly linked to entire gene haplotype 12 (figure 3.4).

Figure 3.4 Mutation network of *CYP1A2* (entire gene) haplotypes observed in the Ethiopian ascertainment and NIEHS populations. Haplotypes from the chimpanzee (Pan troglodytes), orangutan (Pongo pygmaeus) and Rhesus macaque (Macaca mulatta) are shown. Nodes represent haplotypes (shown in table 3.3) and are proportional to haplotype frequencies in the combined Ethiopian ascertainment and NIEHS population (shown in table 3.4). Mutated positions are shown in red along the grey links. All links, except those connecting the chimpanzee, orangutan and Rhesus macaque, are to scale. Recombined haplotypes 22 and 24 were removed from the network to prevent reticulations and are shown in the top right hand corner. Haplotypes are coloured according to their cds haplotype background.

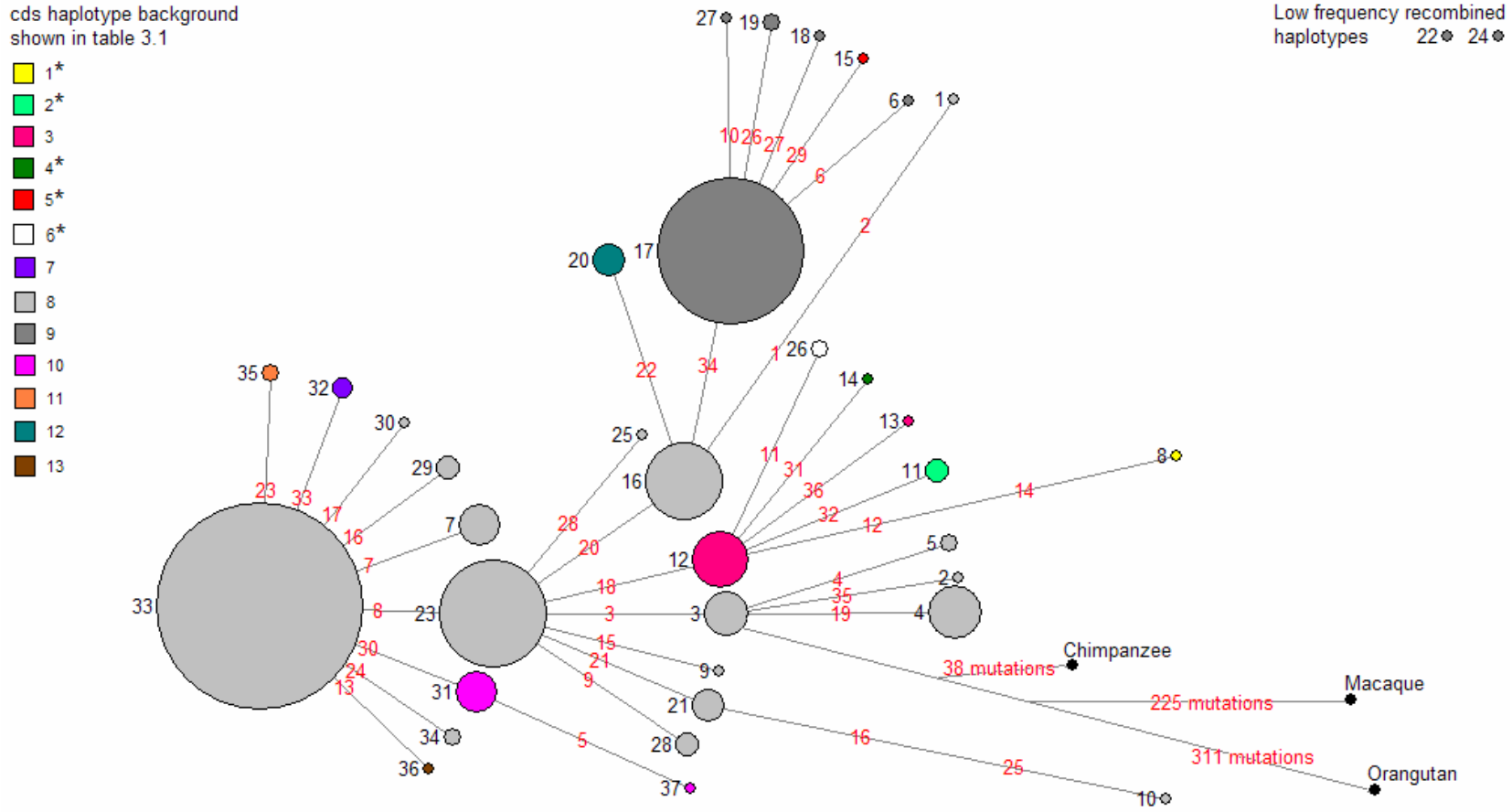


Table 3.3 *CYP1A2* (entire gene) haplotypes used in the mutation networks shown in figure 3.4

Nucleotide change [1]		-1014 C>A	-1008 G>A	-739 T>G	-729 C>T	-592 C>T	-569 G>A	-505 G>A	-163 C>A	-151 G>T	-61 A>G	53 C>G	217 G>A	331 C>T	613 T>G	869 G>C	1352 G>A	1370 G>A	1513 C>A	1589 G>T	2159 G>A	2321 G>C	3463 C>T	3468 A>C	4957 C>G	4961 C>T	5010 C>T	5015 C>G	5029 C>G	5094 T>C	5105 G>A	5253 C>G	5284 C>A	5328 G>A	5347 T>C	5355 G>C	6765 C>T
Mutated position in entire gene phylogenetic network		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Location		5' upstream	5' upstream	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Exon 2	Exon 2	Exon 2	Exon 2	Intron 2	Intron 2	Intron 2	Exon 3	Intron 3	Intron 4	Intron 4	Exon 6	Exon 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	3' UTR	3' UTR
Amino acid change												S18C	G73R	L111F	F205V				S298R				T395M	N397H					F432S	D436N	P485R	Y495Ter	R510Q	N516N		
Haplotype id [2]	1																																			
	2																																			
	3																																			
	4																																			
	5																																			
	6																																			
	7																																			
	8																																			
	9																																			
	10																																			
	11																																			
	12																																			
	13																																			
	14																																			
	15																																			
	16																				A													C		
	17																				A													T		
	18																																			
	19																																			
	20																																			
	21																				G													C		
	22																				G													T		
	23																				G													C		
	24																				G													T		
	25																																			
	26																																			
	27																																			
	28																																			
	29																																			
	30																																			
	31																																			
	32																																			
	33																																			
	34																																			
	35																																			
	36																																			
	37																																			

[1] Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1) from the CYP1A2 genomic reference sequence (NC_000015.8)

[2] White cell, non-derived allele, grey cell, derived allele, nucleotides are also shown to highlight incidences where recombination is inferred

Recombination was inferred between loci 2159 and 5347 in haplotypes 16, 17, 21 and 24 (highlighted in red letters) and between loci 2321 and 5347 in haplotypes 21, 22, 23 and 24 (highlighted in yellow fill)

Table 3.4 *CYP1A2* haplotype (entire gene) frequencies in the world dataset and their distribution among the Ethiopian ascertainment and NIEHS populations.

Haplotype id	Afar		Amhara		Anuak		Maale		Oromo		African American		Yoruba		European		Hispanic		East Asian		World dataset	
	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f	n	f
1	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
2	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
3	3	0.03	2	0.02	5	0.04	0	0.00	4	0.04	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	14	0.020
4	1	0.01	4	0.03	0	0.00	9	0.07	6	0.06	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	20	0.028
5	0	0.00	0	0.00	0	0.00	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.003
6	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.07	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
7	0	0.00	0	0.00	11	0.08	0	0.00	0	0.00	1	0.07	1	0.10	0	0.00	0	0.00	0	0.00	13	0.018
8	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.10	0	0.00	0	0.00	0	0.00	1	0.001
9	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.07	0	0.00	1	0.001
10	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
11	0	0.00	0	0.00	4	0.03	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	4	0.006
12	3	0.03	0	0.00	9	0.06	7	0.05	3	0.03	0	0.00	1	0.10	0	0.00	0	0.00	0	0.00	23	0.033
13	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
14	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
15	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
16	8	0.07	10	0.08	2	0.01	13	0.10	9	0.09	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	42	0.060
17	32	0.27	42	0.34	5	0.04	16	0.12	24	0.24	1	0.07	0	0.00	16	0.67	8	0.57	5	0.19	149	0.212
18	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
19	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.003
20	2	0.02	1	0.01	2	0.01	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	7	0.010
21	1	0.01	1	0.01	3	0.02	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	7	0.010
22	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
23	11	0.09	4	0.03	21	0.15	5	0.04	12	0.12	6	0.43	4	0.40	0	0.00	3	0.21	13	0.50	79	0.112
24	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
25	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
26	0	0.00	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.003
27	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.04	0	0.00	0	0.00	1	0.001
28	0	0.00	2	0.02	0	0.00	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	4	0.006
29	1	0.01	1	0.01	0	0.00	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	4	0.006
30	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
31	0	0.00	1	0.01	1	0.01	9	0.07	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	12	0.017
32	2	0.02	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	3	0.004
33	47	0.40	51	0.41	74	0.53	63	0.48	37	0.36	4	0.29	3	0.30	7	0.29	2	0.14	8	0.31	296	0.420
34	0	0.00	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.003
35	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.003
36	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.07	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
37	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.001
Grand Total	118	1.00	124	1.00	140	1.00	132	1.00	102	1.00	14	1.00	10	1.00	24	1.00	14	1.00	26	1.00	704	1.000

n = number of chromosomes, f = frequency

Recombination was inferred between loci 2159 and 5347 in haplotypes 16, 17, 21 and 24 (highlighted in red letters) and between loci 2321 and 5347 in haplotypes 21, 22, 23 and 24 (highlighted in yellow fill) (table 3.3)

3.3.2 Testing for selection in *CYP1A2*

3.3.2.1 Analysis using statistical tests of neutrality

In order to test the hypothesis that all *CYP1A2* mutations were selectively neutral (Kimura, 1983), three neutrality tests were performed. To control for the effects of different demographic histories, each test was performed on each of the individual Ethiopian ascertainment and NIEHS populations. Since Amhara and Oromo were similar throughout population analyses (see chapter 2), both samples were also combined in an attempt to increase power.

Table 3.5 Results of neutrality tests performed on *CYP1A2* in the Ethiopian ascertainment and NIEHS populations. n = number of chromosomes.

		Tajima's test	McDonald-Kreitman test	Fu and Li's test with an outgroup	
Population	n	Tajima's D (p value)	Fisher's exact test p value	D test (p value)	F test (p value)
Afar	118	-0.88 (p > 0.10)	0.34	-0.84 (p > 0.10)	-1.02 (p > 0.10)
Amhara	124	-1.16 (p > 0.10)	0.32	-4.24 (0.01 < p < 0.02)	-3.68 (0.01 < p < 0.02)
Anuak	140	-1.24 (p > 0.10)	0.32	-1.07 (p > 0.10)	-1.35 (p > 0.10)
Maale	132	-0.86 (p > 0.10)	0.34	-0.88 (p > 0.10)	-1.05 (p > 0.10)
Oromo	102	-0.85 (p > 0.10)	0.34	-2.55 (0.02 < p < 0.05)	-2.30 (0.02 < p < 0.05)
Amhara & Oromo	226	-1.26 (p > 0.10)	0.17	-3.36 (p < 0.02)	-3.03 (p < 0.02)
African American	14	-0.77 (p > 0.10)	1.00	-0.81 (p > 0.10)	-0.95 (p > 0.10)
Yoruba	10	-0.63 (p > 0.10)	0.23	-0.95 (p > 0.10)	-1.03 (p > 0.10)
European	24	0.78 (p > 0.10)	1.00	0.07 (p > 0.10)	0.33 (p > 0.10)
Hispanic	14	-0.53 (p > 0.10)	1.00	0.24 (p > 0.10)	0.38 (p > 0.10)
East Asian	26	0.95 (p > 0.10)	1.00	0.97 (p > 0.10)	1.13 (p > 0.10)

Results of the neutrality tests are shown in table 3.5. Tajima's D was not significantly different from zero in any population ($p > 0.10$ in all cases) and Fisher's exact test p values (two tailed) for each of the McDonald-Kreitman tests were above 0.05. Consequently, the neutral hypothesis was not rejected in each case.

Fu and Li's D and F statistics were not significant at the 5 % significance threshold for all populations except Amhara and Oromo. The negative D and F statistics for Amhara and Oromo were indicative of an excess of recent mutations in the genealogy which is consistent with purifying or positive (not balancing) selection acting on *CYP1A2* (Fu and Li, 1992). Although a Bonferroni correction for multiple tests (11 in this case) suggests that a p value of less than 0.005 would be considered significant, negative D and F test statistics were observed for all Ethiopian ascertainment populations, and Amhara and Oromo were consistent with selection at 5 % significance when analysed separately and together. As a consequence, further analysis was performed in order to try and determine the type of selection. Since *CYP1A2* is highly conserved between species e.g. humans, mice and rats (Aklillu et al., 2003), and non-coding variation is tolerated more than coding variation in humans (see chapter 2), the prior hypothesis was that purifying selection, not positive selection, has been operating on *CYP1A2*.

3.3.2.2 Testing for evidence of purifying selection

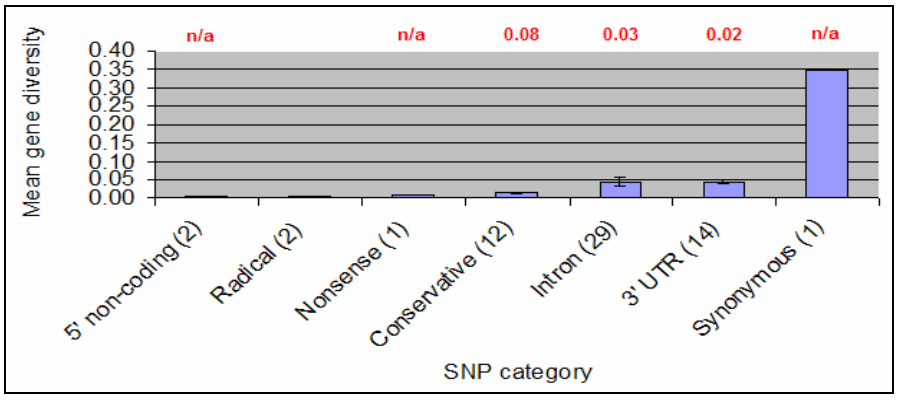
Reduction of both intra-population gene diversity and inter-population genetic distance for non-synonymous SNPs (predicted to cause radical changes to protein structure) in comparison to SNPs in the same genes which have no effect on protein structure is evidence that purifying selection has acted at these non-synonymous SNP sites (Hughes et al., 2003; Hughes et al., 2005). To establish whether purifying selection could be evidenced in *CYP1A2* according to this method, mean gene diversities and genetic distances were computed for variation observed in the Ethiopian ascertainment and NIEHS populations.

3.3.2.2.1 Analysis of intra-population gene diversity

3.3.2.2.1.1 Are reduced gene diversities consistently observed for radical non-synonymous SNPs?

With the exception of the 5' non-coding SNPs, the lowest mean gene diversity (figure 3.5) was observed for radical non-synonymous SNPs. Mean gene diversity was low for nonsense and conservative non-synonymous SNPs, at intermediate levels for intron and 3' UTR SNPs, and highest for the synonymous SNP. In instances where data was sufficiently informative to permit significance tests to be carried out, the difference in mean gene diversity between intron SNPs and radical non-synonymous SNPs was significant as was the difference in mean gene diversity between 3' UTR and radical non-synonymous SNPs (figure 3.5).

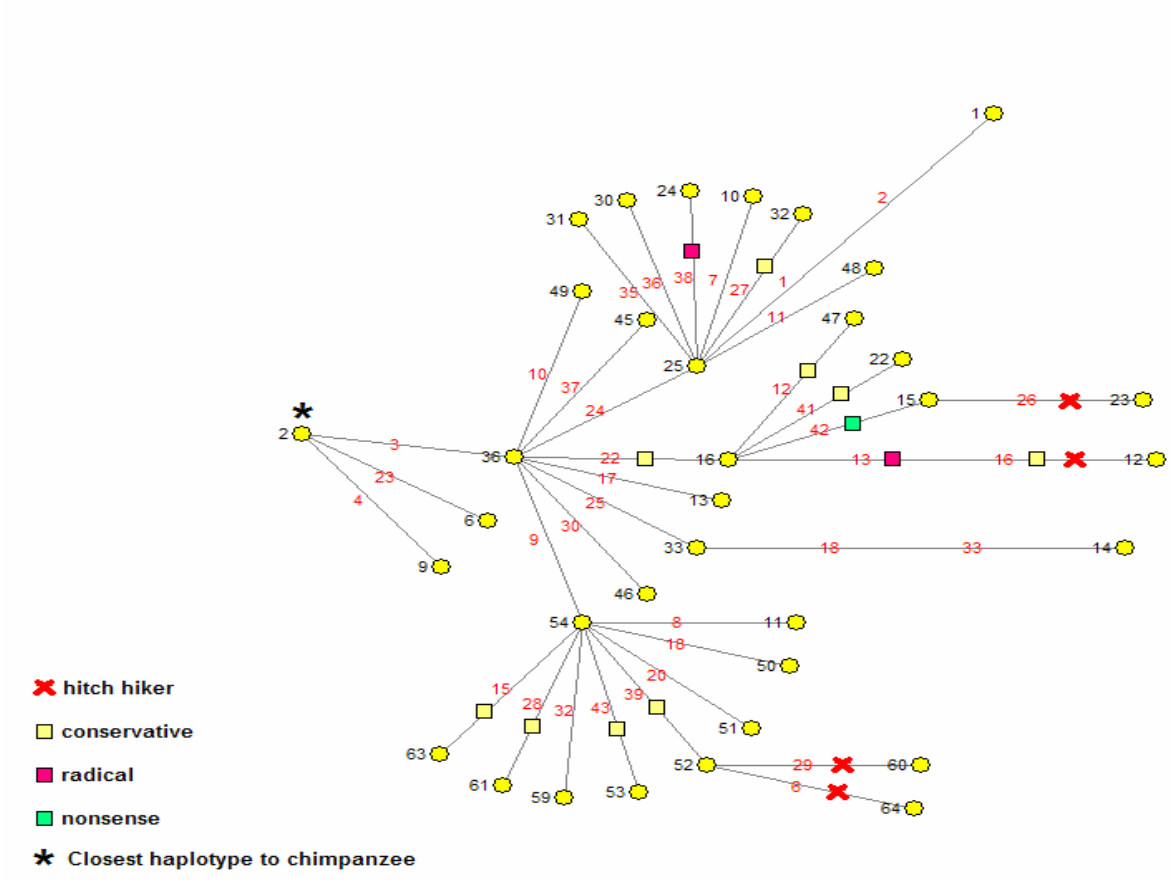
Figure 3.5 Mean gene diversity (heterozygosity) at non-synonymous SNP sites (nonsense, radical or conservative) and synonymous or non-coding sites in *CYP1A2* in the combined Ethiopian ascertainment and NIEHS population. Error bars indicate variance from the mean. Numbers of SNP sites in each category are shown in brackets. One tail p values from t-tests of the hypothesis that mean gene diversity of each SNP category equals that for radical non-synonymous SNP loci are shown in red. n/a = t - test not applicable due to small sample number.



3.3.2.2.1.2 Further analysis in the context of haplotype networks

Particularly low gene diversity values for SNPs occurring on the same haplotype background as non-synonymous SNPs (genetic hitchhikers) might be observed as a result of them being

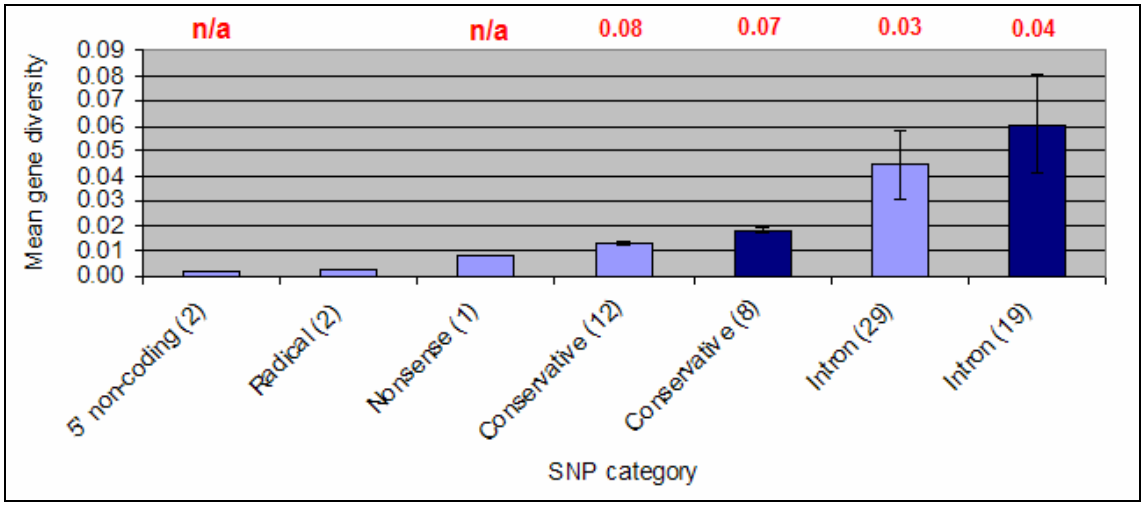
Figure 3.6 Identifying hitchhikers of non-synonymous SNPs from a network of *CYP1A2* haplotypes observed in the Ethiopian ascertainment and NIEHS populations. Nodes represent haplotypes and mutated positions are shown in red along the links. All links are drawn to scale. All 3' UTR SNPs and the single synonymous SNP (5347 C>T in exon 7) were excluded from the analysis to prevent reticulation in the network so that hitchhikers could be easily identified. Non-synonymous SNPs (conservative, radical and nonsense) and their hitchhikers are indicated. Although marker 13 occurred on the background of marker 22, marker 13 was not classified as a hitchhiker (which would have its gene diversity reduced even further as a consequence of being closely linked to a non-synonymous SNP) since it was predicted to be more deleterious than marker 22 (see chapter 2). Intronic markers 6, 26 and 29 were removed from the analysis along with the conservative non-synonymous SNP (16).



Marker in network	CYP1A2 variant	Location
1	-1014C>A	5' upstream
2	-1008G>A	5' upstream
3	-739T>G	Intron 1
4	-729C>T	Intron 1
6	-592C>T	Intron 1
7	-569G>A	Intron 1
8	-505G>A	Intron 1
9	-163C>A	Intron 1
10	-151G>T	Intron 1
11	-61A>G	Intron 1
12	53C>G	Exon 2; S18C
13	217G>A	Exon 2; G73R
15	331C>T	Exon 2; L111F
16	613T>G	Exon 2; F205V
17	869G>C	Intron 2
18	1352G>A	Intron 2
20	1370G>A	Intron 2
22	1513C>A	Exon 3; S298R
23	1589G>T	Intron 3
24	2159G>A	Intron 4
25	2321G>C	Intron 4
26	2534C>T	Intron 5
27	3463C>T	Exon 6; T395M
28	3468A>C	Exon 6; N397H
29	3588G>T	Intron 6
30	3605A>G	Intron 6
32	4957C>G	Intron 6
33	4961C>T	Intron 6
35	5010C>T	Intron 6
36	5015C>G	Intron 6
37	5029C>G	Intron 6
38	5094T>C	Exon 7; F432S
39	5105G>A	Exon 7; D436N
41	5253C>G	Exon 7; P485R
42	5284C>A	Exon 7; Y495Ter
43	5328G>A	Exon 7; R510Q

closely linked to a non-synonymous SNP. To control for this effect, only loci which were not 'hitchhikers of non-synonymous SNPs' were included in the analysis. A network of *CYP1A2* haplotypes was drawn to identify which SNPs were hitchhikers of non-synonymous SNPs (see figure 3.6 for further details). Results were consistent with those observed in the previous analysis except that mean gene diversity values were now higher for SNP categories where genetic hitchhikers of non-synonymous SNPs had been removed (i.e. conservative and intron SNP categories) (figure 3.7).

Figure 3.7 Mean gene diversity (heterozygosity) at various *CYP1A2* SNP sites in the combined Ethiopian ascertainment and NIEHS population once hitchhikers of non-synonymous SNPs were excluded from the analysis. Light and dark blue bars indicate datasets with and without genetic hitchhikers of non-synonymous SNPs, respectively. Error bars indicate variance from the mean. Numbers of SNP sites in each category are shown in brackets. One tail p values from t-tests of the hypothesis that mean gene diversity of each SNP category equals that for radical non-synonymous SNP loci are shown in red. n/a = t - test not applicable due to small sample number.

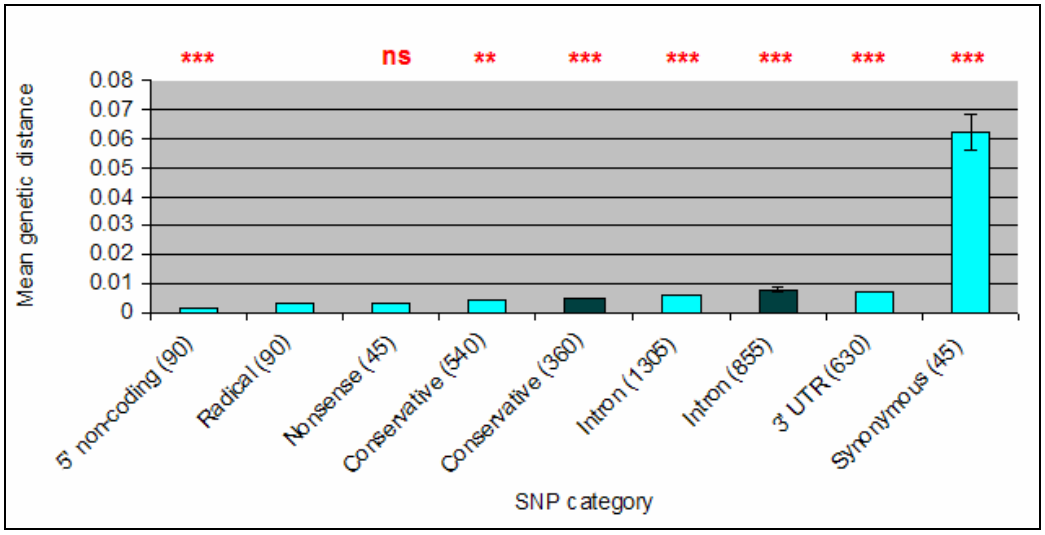


3.3.2.2.2 Analysis of inter-population genetic distance

3.3.2.2.2.1 Are reduced genetic distances consistently observed for radical non-synonymous SNPs?

Consistent with analyses regarding intra-population gene diversity, differences were found among SNP categories with respect to inter-population genetic distance (figure 3.8). With the exception of 5' non-coding SNPs, mean genetic distance was lowest for nonsense SNPs and radical non-synonymous SNPs. Mean genetic distances for conservative, intron, 3' UTR and synonymous SNPs were significantly higher than that for radical non-synonymous SNPs. With the exception of 5' non-coding SNPs, mean genetic distances were higher for SNP categories with no amino acid change than for conservative non-synonymous SNPs. Consistent with the pattern from gene diversity, once non-synonymous hitchhikers were removed from intron and conservative SNP categories (dark green bars in figure 3.8), mean genetic distances between populations increased for these groups.

Figure 3.8 Mean genetic distance values at various *CYP1A2* SNP sites for all inter-population comparisons using individual Ethiopian ascertainment and NIEHS populations. Light green bars indicate datasets using all *CYP1A2* variation. Dark green bars indicate conservative and intron SNP categories from which hitchhikers of non-synonymous SNPs have been removed. Numbers of population comparisons in each SNP category are shown in brackets. One tail p values from t-tests of the hypothesis that mean genetic distance for each SNP category equals that for radical non-synonymous SNP loci are represented as follows: ns = $p > 0.05$, ** = $p < 0.01$, *** = $p < 0.001$



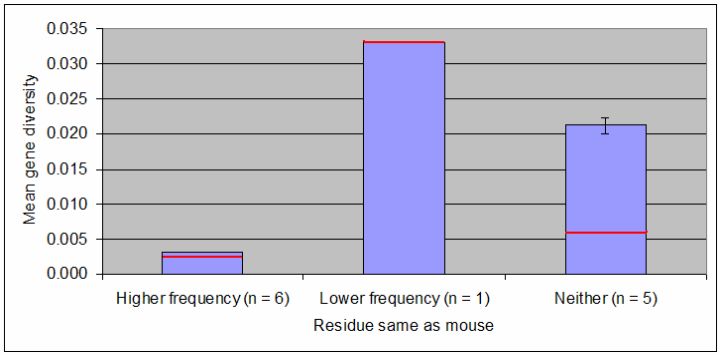
3.3.2.2.3 Evidence of purifying selection acting upon conservative non-synonymous SNPs

To test for evidence of purifying selection on non-synonymous SNPs causing conservative amino acid changes, the *CYP1A2* genomic human reference sequence was compared with the mouse and chimpanzee *CYP1A2* sequences. A mouse and chimpanzee orthologue was available for all twelve of the conservative non-synonymous SNPs observed in the Ethiopian ascertainment and NIEHS populations. At seven (58 %) of these SNP sites, one of the two human alleles encoded an amino acid residue identical to that of the mouse, whereas at five (42 %) of these SNP sites, both human alleles encoded different amino acids to the mouse. At all twelve SNP sites, one of the two human alleles encoded an amino acid residue identical to that of the chimpanzee.

Differences in mean gene diversity were observed among SNP sites in which both amino acids were different from that observed in the mouse, sites at which only the higher frequency residue was different from that observed in the mouse, and sites at which only the lower frequency residue was identical to that observed in the mouse (figure 3.9). Mean gene diversity was lowest (0.003) when the higher frequency amino acid was identical to that observed in the mouse, at an intermediate level (0.021) when neither amino acid was identical to that observed in the mouse, and highest (0.033) when the lower frequency amino acid was identical to that observed in the mouse. Mean (unpaired t test) and median (Kruskal-Wallis test) gene diversities were not however significantly different at the 5 % significance threshold, possibly

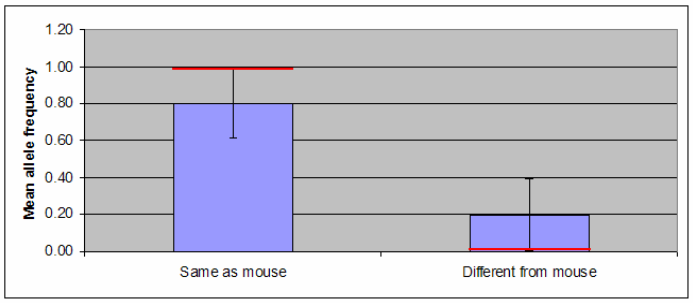
because of the small sample number. These results were in accordance with the hypothesis that, given the functional constraints on evolutionary conserved amino acids, purifying selection acts against mutations which introduce amino acids not observed in the mouse (Hughes et al., 2003).

Figure 3.9 Mean gene diversity at non-synonymous SNP sites causing conservative amino acid changes in the combined Ethiopian ascertainment and NIEHS population. Error bars indicate variances and red horizontal lines indicate median values.



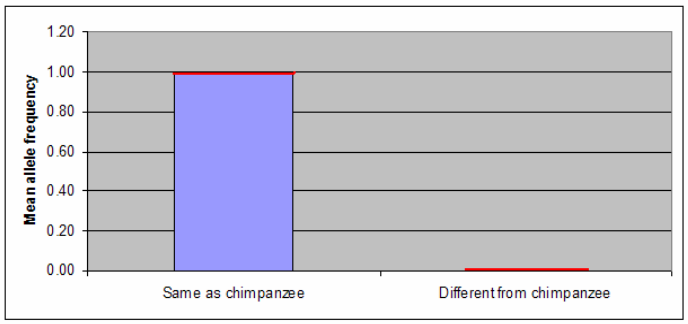
In instances when one of the two human alleles encoded the same amino acid as the mouse, while the other encoded a different amino acid, the mean allele frequency of the former allele (0.802) was four times the mean frequency of the latter allele (0.198) (figure 3.10). The difference between the two mean allele frequencies was not significant ($p > 0.05$ for Wilcoxon matched-pairs signed-ranks test).

Figure 3.10 Mean allele frequency at 12 conservative non-synonymous SNP sites (in the combined Ethiopian ascertainment and NIEHS population) for which a mouse orthologue was available and one SNP encoded a residue identical to the mouse. Error bars indicate variances and red horizontal lines indicate median values.



The mean allele frequency of the allele which encoded a different amino acid to the chimpanzee was less than 1 % (figure 3.11). The difference between the two mean allele frequencies was significant ($p < 0.05$ for Wilcoxon matched-pairs signed-ranks test). Since the higher frequency alleles encode residues which have been conserved within mammals, these results provide further evidence of purifying selection at these conservative non-synonymous SNP sites.

Figure 3.11 Mean allele frequency at 12 conservative non-synonymous SNP sites (in the combined Ethiopian ascertainment and NIEHS population) for which a chimpanzee orthologue was available and one SNP encoded a residue identical to the chimpanzee. Error bars indicate variances and red horizontal lines indicate medians.



3.3.3 *CYP1A2* chronology

3.3.3.1 Estimating the TMRCA of *CYP1A2* allelic variants

The *CYP1A2* sequences and SNPstr genohaplotypes (which incorporated the rs11072507 genotypes with the AC microsatellite haplotypes) were informative enough to date nine *CYP1A2* variants, in addition to the G>C SNP (rs11072507) in the SNPstr, in the Ethiopian ascertainment populations. The majority of variants observed in the Ethiopian ascertainment populations could not be dated because variants were either a) singletons, b) associated with microsatellites which were not variable in size or c) less than eight microsatellite chromosomes were available per variant with the consequence that confidence intervals were so large that dates were uninformative.

3.3.3.2 Coalescent date estimates for *CYP1A2* variants

The unbiased time estimates and associated 95 % confidence intervals for each variant which was dated is shown in table 3.6. Both 2159 G>A and 5347 C>T could not be dated on the background of only the rs11072507 G allele due to small sample numbers. Coalescent date estimates would not however have been significantly different between the rs11072507 G and C background in any case because no more than two G linked chromosomes were available for each variant. In both cases, coalescent dates from rs11072507 G and C combined were used in subsequent analysis.

Consistent with rs11072507 C being the derived allele and rs11072507 G being the ancestral allele (status in the chimpanzee), the coalescent date of rs11072507 C was estimated to be younger than rs11072507 G (table 3.6). The distribution of microsatellite alleles for rs11072507 C and G and for each *CYP1A2* variant dated is shown in figure 3.12. Microsatellite alleles ranged from 16 – 27 and 14 – 25 AC repeat units on the background of rs11072507 C and G

Table 3.6 Inference of the TMRCA (unbiased estimate plus confidence interval) for *CYP1A2* variants and rs11072507. n = chromosome number, G = generations, Y = years, n/a = not applicable. Variants in yellow were assumed to have recombined with rs11072507 (table 3.7) and were consequently dated using microsatellites on the background of each of rs11072507 C and G separately and together (C always produced the younger dates (in blue) which were assumed to be the coalescent dates of the recombination events). All other *CYP1A2* variants (in purple) only occurred on the background of rs11072507 G (table 3.7). Date estimates in green were used in the subsequent analyses. *CYP1A2* variants are arranged in order of increasing TMRCA.

Younger

CYP1A2 variant	Location	Allele dated	rs11072507 background	n	Average Square Distance (ASD)	Time to most common recent ancestor		Equal-tailed 95 % confidence intervals with a star-genealogy model			
								Lower		Upper	
						G	Y	G	Y	G	Y
1589 G>T	Intron 3	T	G	13	0.077	154	4922	22	719	865	27688
2159 G>A	Intron 4	A	C	53	0.925	1849	59168	1160	37133	2999	95955
			G	2	Undetermined						
			G+C combined	55	1.491	2982	95437	1998	63933	4662	149178
5347 C>T	Exon 7	T	C	139	1.626	3252	104058	2493	79773	4253	136109
			G	1	Undetermined						
			G+C combined	140	1.729	3457	110630	2673	85546	4556	145782
rs11072507	SNPstr	C	n/a	270	1.778	3556	113779	2950	94390	4295	137440
5620 A>C	3' UTR	C	C	15	1.533	3067	98131	1414	45245	7262	232394
			G	11	2.182	4364	139635	1906	60992	12693	406176
			G+C combined	26	2	4000	128000	2250	71994	7885	252320
3613 T>C	Intron 6	C	G	24	3.5	7000	224000	4106	131392	14081	450592
1513 C>A	Exon 3	A	G	23	3.826	7652	244870	4403	140896	14785	473120
6324 G>del	3' UTR	-	G	32	5.594	11188	358003	6812	217984	19578	626496
-163 A>C	Intron 1	C	C	23	5.044	10087	322784	5865	187680	19542	625344
			G	184	5.647	11293	361389	9223	295136	14074	450368
			G+C combined	207	5.58	11159	357101	9201	294432	13764	440448
-739 G>T	Intron 1	T	C	250	1.812	3624	115968	2995	95824	4424	141565
			G	380	5.771	11542	369350	9983	319456	13395	428640
			G+C combined	630	4.2	8400	268800	7517	240541	9439	302048
rs11072507	SNPstr	G	n/a	450	5.982	11964	382861	10450	334400	13793	441376

Older

respectively. The modal microsatellite allele size was 23 AC repeats for both rs11072507 C and G, and for the majority of *CYP1A2* variants (figure 3.12).

Figure 3.12 Distributions of AC microsatellite alleles which were used to date rs11072507 (SNPstr SNP) and various *CYP1A2* variants observed in the Ethiopian ascertainment populations

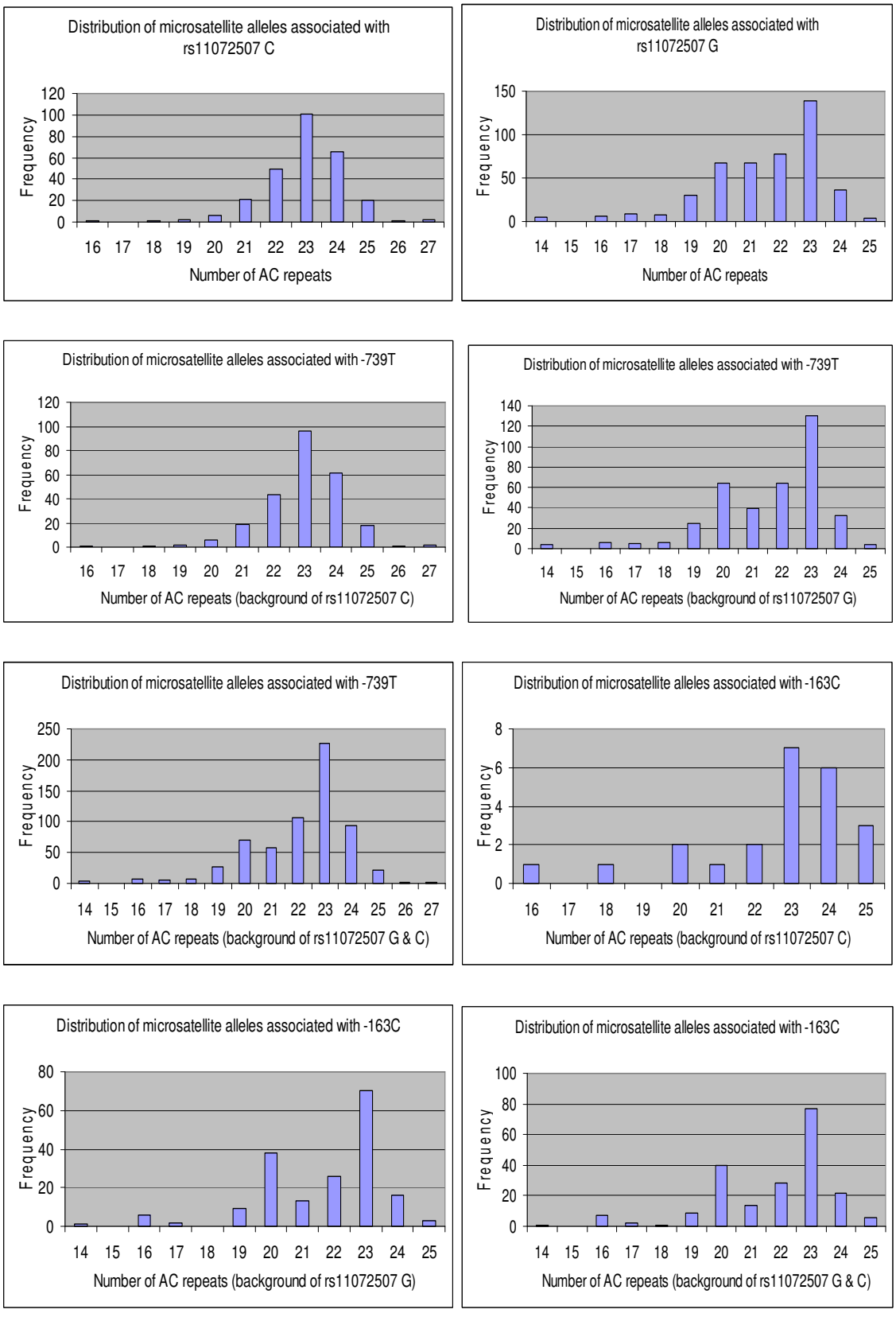
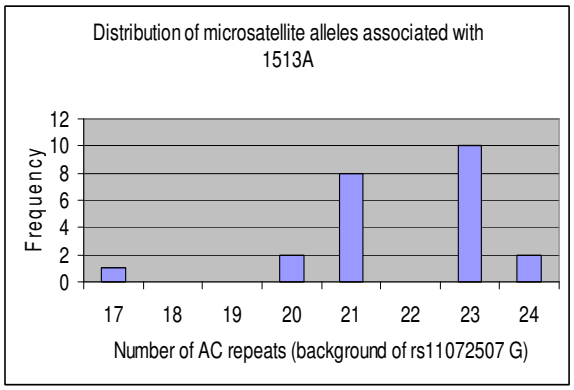
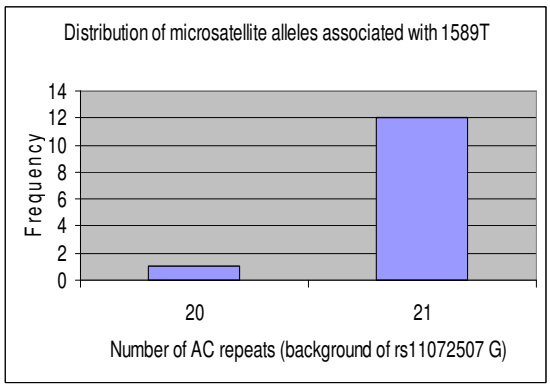
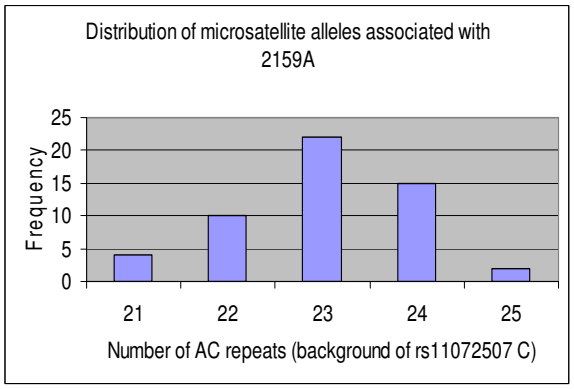
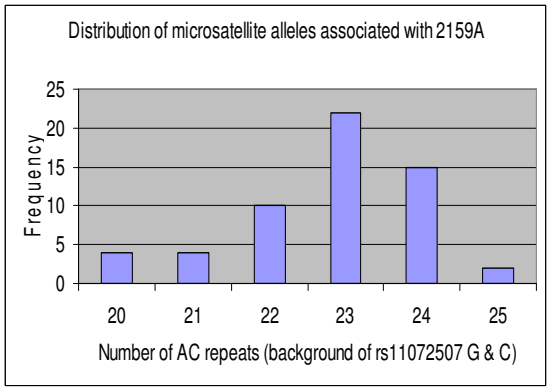


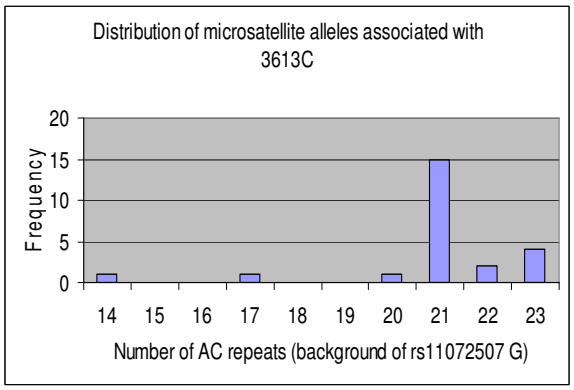
Figure 3.12 continued

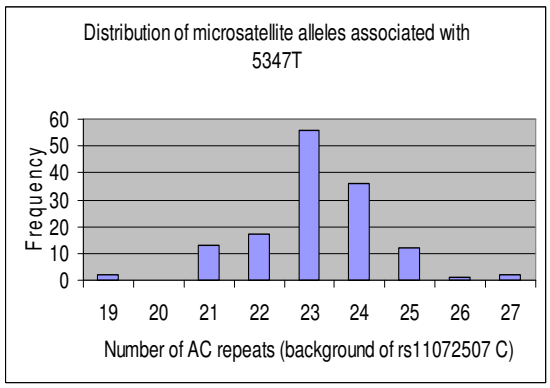


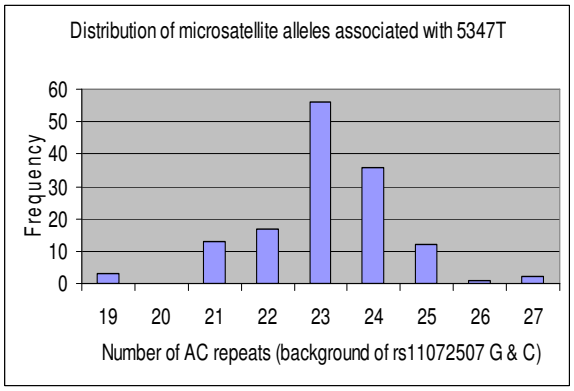












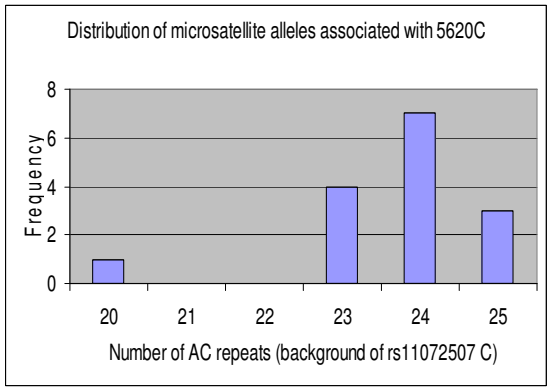
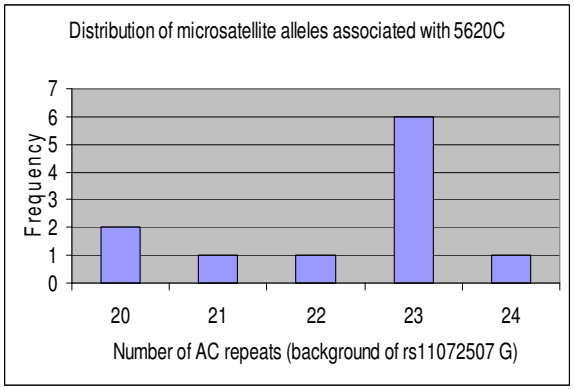
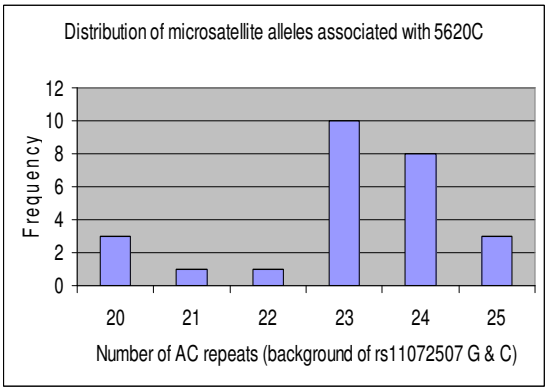
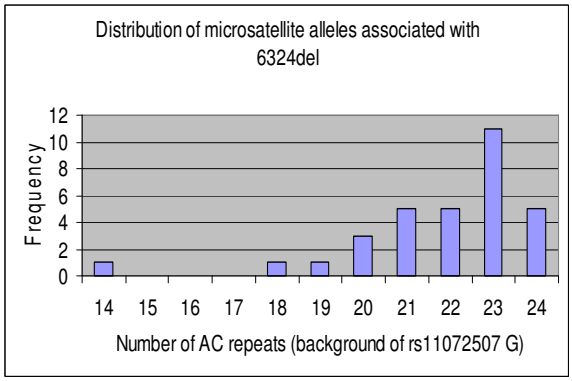


Figure 3.12 continued







3.3.3.3 Are the dates consistent with each other?

Of all the coalescent dates, the date for rs11072507 G was estimated to be the oldest (table 3.6). This was expected given that rs11072507 G was observed in the chimpanzee (ancestral allele) whereas all of the dated *CYP1A2* variants were not (i.e. they were derived alleles) and would have post-dated rs11072507 G.

All coalescent dates estimated on the background of either rs11072507 C or G did not pre-date the rs11072507 C or G allele respectively, with one exception. The coalescent date of -163 A>C, dated on the background of rs11072507 C, was estimated to be older than rs11072507 C (table 3.6). This can however be due to the recombination event involving both -163 A>C and rs11072507 (table 3.7).

In order to establish whether the date estimates were consistent with the evolution of *CYP1A2* in humans, a rooted network of haplotypes (*CYP1A2* entire gene plus rs11072507 SNP) observed in the Ethiopian ascertainment population was produced. The network is shown in figure 3.13 whilst the haplotypes are shown in table 3.7. All SNP coalescence date estimates (table 3.6) were consistent with the relative positions of the SNPs in the mutation network (figure 3.13), i.e. SNPs on internal branches were older than SNPs on external branches.

Table 3.7 *CYP1A2* plus rs11072507 haplotypes. Haplotype network is shown in figure 3.13.

Nucleotide change [1]		-1014 C>A	-1008 G>A	**-739 G>T**	-729 C>T	-592 C>T	-505 G>A	**-163 A>C**	-151 G>T	53 C>G	1352 G>A	1370 G>A	**1513 C>A**	**1589 G>T**	**2159 G>A**	2321 G>C	3463 C>T	3468 A>C	4957 C>G	4961 C>T	5010 C>T	5015 C>G	5094 T>C	5105 G>A	5253 C>G	5284 C>A	5328 G>A	**5347 C>T**	5355 G>C	6765 C>T	**rs11072507 G>C**
Mutated position in mutation network		1	2	3	4	5	7	8	9	11	16	17	18	19	20	21	22	23	24	25	26	27	29	30	31	32	33	34	35	36	37
Location		5' upstream	5' upstream	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Exon 2	Intron 2	Intron 2	Exon 3	Intron 3	Intron 4	Intron 4	Exon 6	Exon 6	Intron 6	Intron 6	Intron 6	Intron 6	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	3' UTR	3' UTR	SNPstr
Amino acid change										S18C			S298R				T395M	N397H					F432S	D436N	P485R	Y495Ter	R510Q	N516N			
Haplotype id [2]	1																														
	2																														
	3																														
	4																														
	5																														
	6																														
	7																														
	8																														
	9																														
	10																														
	11																														
	12																														
	13																														
	14																														
	15																														
	16																														
	17																														
	18																														
	19																														
	20																														
	21																														
	22																														
	23																														
	24																														
	25																														
	26																														
	27																														
	28																														
	29																														
	30																														
	31																														
	32																														
	33																														
	34																														
	35																														
	36																														
	37																														

[1] Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1) from the *CYP1A2* genomic reference sequence (NC_000015.8)

[2] White cell, non-derived allele, grey cell, derived allele.

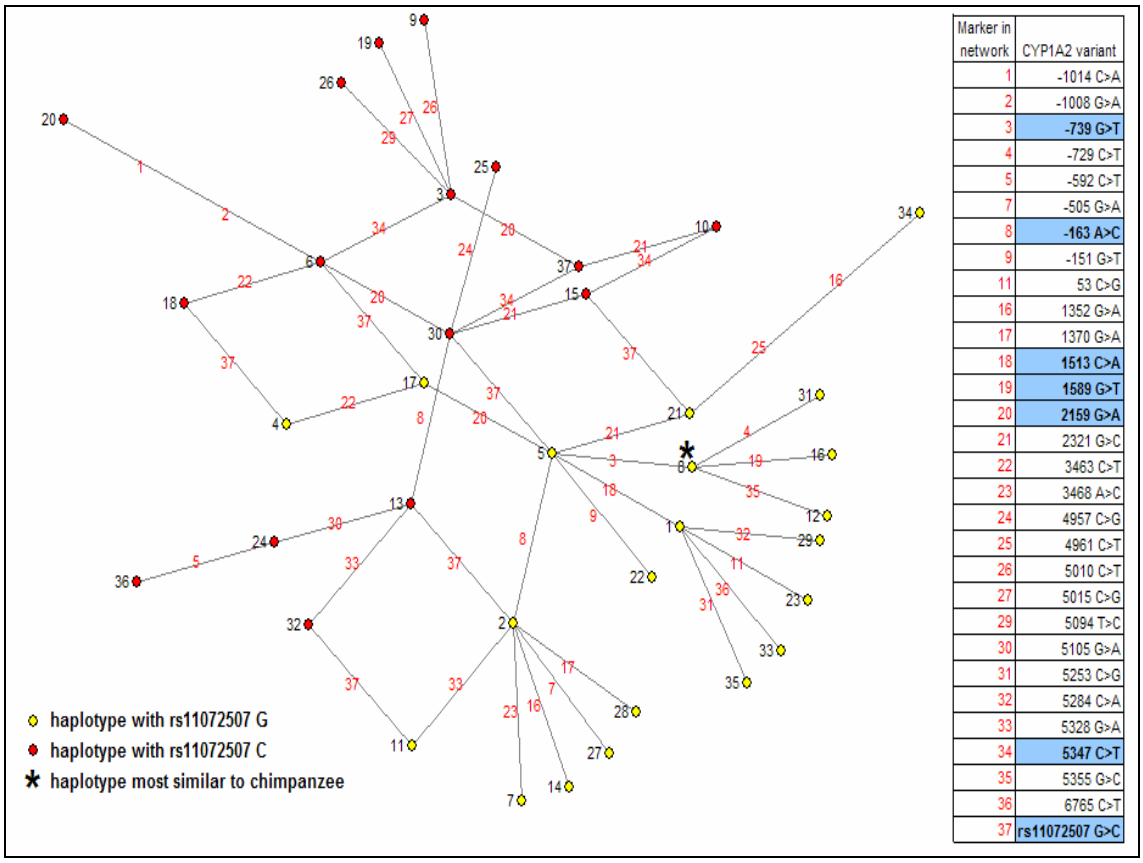
Nucleotide changes in bold were dated. 3613 T>C (intron 6), 5620 A>C (3' UTR) and 6324 G>del (3' UTR) were also dated but are not shown here because their allelic status in the chimpanzee and macaque is not known and haplotypes shown here were used to construct a mutation network (figure 3.13). Haplotype analysis did however reveal that 3613 T>C and 6324 G>del only occurred on the background of rs11072507 G, whilst 5620 A>C occurred on the background of both rs11072507 G and C (data not shown).

SNPs highlighted in purple only occurred on the background of rs11072507 G.

SNPs highlighted in yellow were observed on the background of both rs11072507 G and C.

Although recombination was not observed between -739 G>T (highlighted in green) and rs11072507 (only three haplotypes observed between the two SNPs), since -163 A>C only occurred on background of -739 T and since recombination was observed between -163 A>C and rs11072507 (four haplotypes observed between the two SNPs) and -739 T and -163 C are located close to each other, recombination was assumed to have occurred on the background of -739 T.

Figure 3.13 Mutation network of haplotypes (*CYP1A2* entire gene plus rs11072507 SNP) observed in the Ethiopian ascertainment populations. Nodes represent haplotypes (table 3.7) and mutated positions are shown in red along the grey links. All links are drawn to scale. The network was rooted using haplotypes from the chimpanzee and macaque. Dated SNPs are highlighted in blue. Alleles at positions 3613, 5620 and 6324 (which have been dated) could not be included in the network since their status in the chimpanzee and macaque was not known. The network is consistent with the network shown in figure 3.4.



Marker in network	CYP1A2 variant
1	-1014 C>A
2	-1008 G>A
3	**-739 G>T**
4	-729 C>T
5	-592 C>T
7	-505 G>A
8	**-163 A>C**
9	-151 G>T
11	53 C>G
16	1352 G>A
17	1370 G>A
18	**1513 C>A**
19	**1589 G>T**
20	**2159 G>A**
21	2321 G>C
22	3463 C>T
23	3468 A>C
24	4957 C>G
25	4961 C>T
26	5010 C>T
27	5015 C>G
29	5094 T>C
30	5105 G>A
31	5253 C>G
32	5284 C>A
33	5328 G>A
34	**5347 C>T**
35	5355 G>C
36	6765 C>T
37	**rs11072507 G>C**

3.3.3.4 Approximate dates for non-synonymous *CYP1A2* variants

Dates could not be estimated from microsatellite data for many of the non-synonymous variants observed in the Ethiopian ascertainment populations. The positions of these variants within the network (figure 3.13), relative to dated variants, were however used to estimate their coalescent dates as follows:

- Since 53 C>G (S18C at mutated position **11**), 5253 C>G (P485R at mutated position **31**) and 5284 C>A (Y495Ter at mutated position **32**) occurred on the background of 1513 C>A (at mutated position **18**), all three variants were expected to be younger than 1513 C>A (245,000 years).

- Since 3468 A>C, (N397H at mutated position **23**) and 5105 G>A (D436N at mutated position **30**) occurred on the background of rs11072507 G>C (at mutated position **37**), both variants were expected to be younger than rs11072507 G>C (114,000 years).

- Since 3463 C>T (T395M at mutated position **22**) and 5094 T>C (F432S at mutated position **29**) were on the background of 2159 G>A (at mutated position **20**), both variants were also expected to be younger than 2159 G>A (95,400 years).

3.4 Discussion

3.4.1 Network analysis of *CYP1A2* haplotypes

This study has unveiled a network of *CYP1A2* haplotypes observed in the Ethiopian ascertainment and NIEHS populations and has compared human *CYP1A2* DNA sequences to those found in chimpanzee, orangutan and Rhesus macaque genomes. Of all the *CYP1A2* exons, exon 5 was the most conserved throughout primate evolution showing 100 % similarity to the chimpanzee's sequence. Exons 4 and 1 were also highly conserved among the primates. These results are unsurprising since exon 1 is the 5' UTR and contains the leader sequence of the gene (Ikeya et al., 1989), thus playing a crucial role in protein expression, and exon 4 is thought to contain part of the enzyme's active site (Sansen et al., 2007). Situated in the middle of the gene and in close proximity to exon 4, exon 5 is likely to play a pivotal role in the structure and functioning of the protein. Exon 2 on the other hand was the least conserved exon. This too is expected given its distance from the location of the presumed active site.

Consistent with other studies investigating relationships between primates (e.g. Bailey et al., 1991), the chimpanzee, orangutan and macaque were generally linked to the human clade via a single common branch with the chimpanzee being the most similar to the humans, usually followed by the orangutan, then the macaque. In some networks, the orangutan was the most dissimilar to the humans, however this may be explained by missing data since the orangutan *CYP1A2* sequence has yet to be fully characterised.

Contrary to the combined cds region, the modal human haplotype for the entire *CYP1A2* gene, which was observed in all Ethiopian ascertainment and NIEHS populations, was not the most similar human node to the chimpanzee. Consequently, it appears that non-coding *CYP1A2* variants have been tolerated more than coding *CYP1A2* variants throughout evolution, a finding which is to be expected if purifying selection has been taking place. A haplotype confined to Ethiopia, and which was connected to many other haplotypes specific to Ethiopia, was most similar to the chimpanzee in the *CYP1A2* entire gene network. In light of this, it appears that out of the populations included in this analysis, the Ethiopians may have retained ancestral forms of *CYP1A2*.

In the vast majority of cases, single mutations (single steps) linked human *CYP1A2* haplotypes together. Retention of almost all the steps between haplotypes is consistent with population expansion. In some instances, intermediate haplotypes were missing in the networks.

Potentially common intermediate haplotypes were missing in the NIEHS sample sets (since intermediate haplotypes linking common haplotypes together were not observed), whilst potentially rare intermediate haplotypes were missing in the Ethiopians (since intermediate haplotypes connecting two singletons to their respective progenitor haplotypes were not found). These findings are likely to reflect the power of the sample sizes used in this study to detect variants within the actual populations (refer to the discussion in chapter 2 for further information regarding sample sizes and power calculations).

All of the *CYP1A2* haplotypes predicted to code for a protein with altered activity were observed in the external branches of the networks, and may consequently be recent (coalescent dates are discussed towards the end of this chapter). They appeared to evolve from two haplotypes and the vast majority evolved from just one (which is characterised as entire gene haplotype 12 in this chapter). In view of this, perhaps entire gene haplotype 12 is more mutable than others. The haplotype itself is predicted to code for an undamaged *CYP1A2* protein, but could be in linkage disequilibrium with a variant outside of the gene that increases its susceptibility to mutations likely to alter the structure/function of the protein.

3.4.2 Testing for selection in *CYP1A2*

Fu and Li's tests pointed towards selection (purifying or positive) in *CYP1A2* in Amhara and Oromo but Tajima's D and the McDonald-Kreitman test did not detect selection in any of the populations analysed in this study. These neutrality test statistics do not have the ability to distinguish between demographic effects and selection, and lack power to detect very recent selection pressures when the selected lineage may be in the minority. The methods also have little power to detect selection when intragenic recombination (Wall, 1999) and population structuring are evident (Przeworski, 2002). Given that all of the *CYP1A2* haplotypes predicted to code for a protein with altered activity may be recent, and recombination was inferred in the Ethiopian ascertainment populations, and since it is common place for Ethiopian populations to form hierarchical societies (Freeman & Pankhurst, 2003), it is therefore possible that selective pressures operating on *CYP1A2* would not be detected by these commonly used neutrality tests.

A study involving 2784 SNPs, in 396 protein coding genes, in an ethnically diverse sample showed that reduction of both intra-population gene diversity and inter-population genetic distance for non-synonymous SNPs (predicted to cause radical changes to protein structure) in comparison to SNPs in the same genes which have no effect on protein structure is evidence that purifying selection has acted at these non-synonymous SNP sites (Hughes et al., 2003; Hughes et al., 2005). When this approach was applied to the variation observed in *CYP1A2* in the Ethiopian ascertainment and NIEHS populations, intra-population gene diversities and inter-population genetic distances were generally lowest for SNPs expected to have greatest impact on protein structure (radical non-synonymous SNPs and the nonsense mutation in exon 7).

These results are consistent with the hypothesis that purifying selection has affected the allele frequencies of SNPs, predicted to have greatest impact on protein structure, in *CYP1A2* in humans. Evidence of purifying selection at these SNP sites was strengthened when genetic hitchhikers of non-synonymous SNPs (which may affect differences in gene diversity and genetic distance) were taken into account. This variable was not considered by Hughes et al. (2003) and Hughes et al. (2005), possibly due to the nature of the dataset studied.

Purifying selection was also evidenced in the case of conservative non-synonymous SNPs, as it was in Hughes et al. (2003), through comparison of human sequences with orthologous mouse sequences at conservative non-synonymous SNP sites. Comparisons with chimpanzee sequences were also performed to increase the depth of the analysis. Results were in accordance with the hypothesis that, given the functional constraints on evolutionary conserved amino acids, purifying selection acts against mutations which introduce amino acids not observed in the mouse or chimpanzee. Further evidence of purifying selection at these conservative non-synonymous SNP sites was also provided by the observation that the higher frequency human alleles, at all conservative SNP sites, encoded residues which have been conserved throughout mammalian evolution.

Contrary to our expectations, gene diversity and genetic distance for 5' non-coding SNPs was often lower than that for radical non-synonymous SNPs. This may be explained by small sample size since there were only two SNPs in the 5' non-coding category. It is possible that the results would be different if the numbers were increased. It is noteworthy to mention that variability in only one gene was analysed in this study, as opposed to 396 in Hughes et al. (2005). Although this study may have lacked power in the number of genes analysed, it has shown that Ethiopians are a good population for detecting signatures of purifying selection in humans because, through being an old population and maintaining a relatively large population size, they have retained a number of mutations likely to be under purifying selection.

Contrary to expectations for a SNP which introduces a premature stop codon into the gene, mean gene diversites and genetic distances for the nonsense SNP were never lower than those for radical non-synonymous SNPs. In view of this, it is possible that this nonsense SNP may not be too important to the structure/function of the enzyme given its proximity to the end of the gene (exon 7).

Population frequencies of the minor allele at SNP sites which showed evidence of purifying selection were generally in the range of 1-10 %. Other studies (Hughes et al., 2003; Hughes et al., 2005; Wong et al., 2003) have also reported numerous non-synonymous SNPs with similar minor allele frequencies in human genes. These frequencies are substantially higher than those of genes causing severe Mendelian disease phenotypes such as Huntington's Chorea and cystic fibrosis (McKusick and Francomano, 1997), suggesting that these non-synonymous SNPs are not as deleterious as those causative of severe disease types. This in turn suggests that the selective forces acting against these non-synonymous SNPs are modest in comparison

to those at SNP sites causative of severe disease (Hughes et al., 2003; Hughes et al., 2005). Since mutations associated with complex diseases are expected to be individually only slightly deleterious, as opposed to highly deleterious variants associated with Mendelian diseases, evidence of mild purifying selection may be used to identify candidate alleles for complex disease-association studies (Hughes et al., 2003).

As reviewed in Hughes et al. (2003) and Hughes et al. (2005), the observation of purifying selection acting against slightly deleterious alleles which are present at relatively high frequencies in the human population is in accordance with the nearly neutral theory of molecular evolution (Ohta, 1973; Ohta, 1976; Ohta, 2002). This theory predicts that slightly deleterious mutations can reach high frequencies (effectively acting as if they were selectively neutral) due to genetic drift when population size is small. As effective population size increases and purifying selection becomes more effective, it is predicted that slightly deleterious alleles, which reached high frequencies in bottlenecked populations due to drift, will decrease gradually over time. It is a widely accepted view that the human population has expanded since the origin of anatomically modern human 100,000 – 250,000 years ago and has experienced a number of bottlenecks (Harpending et al., 1998; Watkins et al., 2003; Zhitovsky et al., 2003). The mutations evidenced to be under purifying selection in this study may be recent in origin, but may also include variants which drifted to high frequencies in smaller ancestral populations.

3.4.3 *CYP1A2* chronology

Human and chimpanzee lineages are generally assumed to have diverged approximately 5-7 million years ago (Chen and Li, 2001; Carroll, 2003) with the human genus (*Homo*) originating in Africa (Horai et al., 1995; Ingman et al., 2000; Thomson et al., 2000; Ke et al., 2001). Morphological traits from fossils support the idea of a recent origin for modern humans in Africa less than 200,000 years ago (White et al., 2003; McDougall and Fleagle, 2005; Campbell and Tishkoff, 2008). According to the 'Out of Africa' model of human evolution, modern humans originated in Africa and colonised the rest of the globe in the last ~ 100,000 years (Tishkoff and Verrelli, 2003). Evidence suggests that humans may have migrated out of Africa, reaching the Middle East and Europe, Asia and Australia by 30 – 50,000 years ago, and the Americas some 15 – 30,000 years ago (Campbell and Tishkoff, 2008).

The coalescent date estimates of the *CYP1A2* variants in this study were old in terms of the emergence of anatomically modern human and all, except for 1589 G>T, pre-dated the expansion of humans out of Africa, less than ~ 100,000 years ago. In fact, five variants (-739 G>T and -163 A>C, both of which are in intron 1, 1513 C>A in exon 3 causing S298R, 3613 T>C in intron 6 and 6324 G>del in the 3' UTR) were estimated to have arisen prior to the emergence of modern humans in Africa. We saw from chapter 2 that whilst 1589 G>T (the youngest *CYP1A2* variant in this study) was confined to Ethiopia, -163 C>A and -739 T in intron 1, 2159 G>A (intron 4) and 3613 T>C (intron 6) were found in populations worldwide. This is

not unexpected given their relative coalescent date estimates. On the other hand however, 1513 C>A (S298R) and 6324 G>del (3' UTR) were only found at appreciable frequencies in populations with a recent African ancestry and may belong to a subset of alleles which did not leave Africa. Likewise, 5620 A>C (3' UTR) was only observed in Ethiopian populations and despite its age may also belong to a subset of alleles which did not leave Ethiopia. Interestingly, non-synonymous 5347 C>T (exon 7) was observed on a global scale but was not found in Yoruba and may have not therefore been transferred from East to West Africa during the early expansion of African populations. The G>C SNP (rs11072507) in the SNPstr has been reported in the Japanese (14.6 %) and Chinese (10.5 %) HapMap populations (http://www.hapmap.org/) but was also not observed in Yoruba. Perhaps this allele also followed a similar route to 5347 C>T.

Although many of the coalescent dates for non-synonymous variants could not be estimated using microsatellite data, they could be placed within a time frame because the positions of these non-synonymous variants in the *CYP1A2* network, relative to variants which were dated, were known. The approximate time frame of these alleles was consistent with the hypothesis that the mutations evidenced to be under purifying selection in the Ethiopian ascertainment and NIEHS populations may include variants which drifted to high frequencies in smaller ancestral populations.

Although the coalescent date estimates appear to be consistent with the human *CYP1A2* mutation network, they should be treated cautiously. Their confidence intervals are large and more samples and/or microsatellites are needed to improve time point estimates and reduce the confidence intervals. Furthermore, a novel A>C SNP was observed in the Amhara (8.3 %) at the penultimate base of the AC microsatellite which was used to date *CYP1A2* alleles in this study. This variant may affect the mutation rate of the microsatellite which is important in determining estimates of the time to the most recent common ancestor.

3.5 Conclusion

This study has produced a rooted *CYP1A2* haplotype network which is consistent with humans evolving from non-human primates with the chimpanzee being their closest living relative. Of the populations used in the network, which included those with a recent African ancestry, Europeans, Hispanics (partial African ancestry) and East Asians, a *CYP1A2* haplotype in the Ethiopian dataset was most similar to that of the chimpanzee. Haplotype networks revealed a varying level of conservation among the *CYP1A2* exons and showed that all haplotypes predicted to code for a protein with altered activity were observed in the external branches of the network. Furthermore, most of these potentially damaging haplotypes evolved from a single haplotype which may be susceptible to non-synonymous mutations. Commonly used neutrality tests lacked power in detecting signatures of selective processes in this study. Purifying selection was however evidenced in *CYP1A2* in the human population through an approach

which takes into account intra-population gene diversity and inter-population genetic distance. The time to most common recent ancestor of nine *CYP1A2* variants was estimated using an AC microsatellite situated 5.6 kb downstream of the end of *CYP1A2*. Coalescent date estimates place most variants into a period which pre-dates the expansion of anatomically modern human out of Africa and into the New World, e.g. -739 G>T in intron 1 is estimated to be ~369,000 years old with a 95 % confidence interval of 319,000 – 429,000 years.

---

4 Can data reported by the CYP450 Allele Nomenclature Committee be used to design a diagnostic test to predict CYP1A2 functional variation in Ethiopian populations?

4.1 Introduction

Individualised medicine is a concept in which healthcare intervention utilises information about a patient's specific characteristics (including genes, phenotype and lifestyle) to better match intervention to an individual's needs (Kawamoto et al., 2009). Such information could be used to help distinguish disease status, choose between different medications and/or tailor their dosage, provide a specific therapy, or initiate a preventative measure at an appropriate time. Individualised medicine is not in widespread clinical use, but some aspects of the approach have been established in medical practice (Arnett et al., 2009). As an example, knowledge of polymorphisms in *CYP2C9* and vitamin K epoxide reductase (*VKORC1*), has led to commercially available tests which enable more accurate dosing of the anticoagulant drug warfarin (http://www.fda.gov/bbs/topics/NEWS/2007/NEW01701.html). The tests are based on algorithms that take into account the age, gender, weight, and *CYP2C9* and *VKORC1* genotypes of an individual.

Individualised medicine may be unrealistic in Africa for the foreseeable future due to sparse funds and rudimentary healthcare infrastructure. Since CYP450 allelic variants are known to be distributed differently among ethnic groups and/or geographic regions (see chapter 2 with respect to *CYP1A2*), it may be feasible to decide between alternative medical intervention options using group derived pharmacogenetic profiles when individual characterisation is not possible. An aim of the research programme of which this project is part, is to develop a diagnostic test (based on genetic markers) to predict the functional variation of CYP1A2 among different Ethiopian populations, with a view to seeing if ethnic identity, or geographic location, is useful in predicting efficacy and safety of therapeutic interventions. The use of ethnicity or geographic location as a proxy for the prediction of drug response could potentially be beneficial to medical practice and public health policy makers in Ethiopia. For instance, medical professionals may be able to prescribe drugs most likely to be most effective in the group being treated, when the genetic profiles of individuals are not known.

The diagnostic test will be based on genetic markers, within *CYP1A2*, which should offer a higher throughput, more practical and efficient approach than phenotype testing would (e.g. with caffeine).

The importance of predicting CYP1A2 functional variation in Ethiopian populations is highlighted by the fact that many pharmaceuticals, which may be metabolised by CYP1A2, are

administered in Ethiopia. Examples include primaquine and praziquantel (Li et al., 2003) which are used as the first line of treatment for malaria and schistosomiasis respectively (Federal Democratic Republic of Ethiopia Ministry of Health, 2004). Other examples of pharmaceuticals metabolised by CYP1A2 are shown in table 4.1.

Table 4.1 Pharmaceuticals metabolised by CYP1A2 (derived from http://medicine.iupui.edu/flockhart/table.htm)

Drug	Treatment
Phenacetin	Pain relief
Naproxen	
Acetaminophen	
Mexiletine	Antiarrhythmic
Warfarin	Anticoagulant
Amitriptyline	Antidepressant
Clomipramine	
Fluvoxamine	
Imipramine N-DeMe	
Ondansetron	Antiemetic
Clozapine	Antipsychotic
Haloperidol	
Olanzapine	
Propranolol	Hypertension
Verapamil	Hypertension, angina, certain heart rhythm disorders
Zileuton	Respiratory disease
Ropivacaine	Local anaesthetic
Theophylline	Respiratory disease
Cyclobenzaprine	Muscle relaxant
Tizanidine	
Tacrine	Alzheimer's disease
Zolmitriptan	Migraine
Estradiol	Menopause or prostate cancer, ovarian failure, hypogonadism
Riluzole	Motor neuron disease

4.1.1 Aim

The aim of this project is to establish whether it is possible to predict population level CYP1A2 functional variation in Ethiopia on the basis of known variants. The following steps are required in order to do this:

Step 1: Design a typing strategy, based on previously reported sequence and functional data, to predict CYP1A2 functional variation among populations. Begin by identifying all known *CYP1A2* allelic variants reported by the CYP450 Allele Nomenclature Committee and assign each variant with an associated phenotype from previously reported functional data. The predicted phenotype categories may be: non-functional, reduced metabolic activity, increased metabolic activity, normal (CYP1A2*1A-like) metabolic activity or unknown function. Generate a network of these *CYP1A2* alleles to help select the minimum number of polymorphisms which need to be typed in order to predict CYP1A2 phenotypes. Compile an algorithm to determine the predicted CYP1A2 phenotype based on *CYP1A2* polymorphic states.

Step 2: Establish what CYP1A2 functional variation is predicted to exist in Ethiopia by sequencing all seven *CYP1A2* exons and flanking introns in the Ethiopian ascertainment populations.

Step 3: Apply the algorithm from step 1 to variation observed from step 2 to establish the extent to which the algorithm, based on known variation, predicts functional variation predicted from *CYP1A2* sequences in the Ethiopian ascertainment populations.

Step 4: If the algorithm based on known variation fails to predict, or incorrectly predicts CYP1A2 functional variation predicted from sequencing the Ethiopian ascertainment populations, establish how much genotype/phenotype testing is required to improve the diagnostic test for Ethiopia.

4.2 Methods

The Ethiopian ascertainment sequences described in chapter 2 were used in this study in addition to genotype data from -3860 G>A and -2467 delT (see below). Missing data was excluded from the analysis and haplotypes were inferred by the ELB approach (Excoffier et al., 2003). Refer to the methods section in chapter 2 for details regarding sample collection, DNA extraction, sequencing, testing for deviation from Hardy Weinberg equilibrium, haplotype inference and CYP1A2 phenotype predictions.

Mutation networks were constructed using Network software, version 4.510 (fluxus-engineering.com). Median joining networks were constructed to limit levels of reticulation (Bandelt et al., 1999). The algorithm used to construct these median joining networks is based on the limited introduction of likely ancestral sequences/haplotypes into a minimum spanning network of the observed sequences. The resultant networks were drawn using Network publisher, version 1.1.0.7 (Fluxus Technology Ltd).

4.2.1 Genotyping of -3860 G>A and -2467 T>-

Genotyping of -3860 G>A (rs35694136) and -2467 T>- (rs2069514) was performed using the ABI TaqMan SNP genotyping assay C__60142977_10 and C__15859191_30 respectively (Applied Biosystems (ABI), Warrington UK). Each assay contained a mix of unlabeled PCR primers and TaqMan MGB probes (FAM and VIC dye-labeled) for the allelic discrimination of each SNP. Details of the oligonucleotides are not disclosed and are patented by ABI (US patents and corresponding patent claims outside the US: 5,538,848, 5,723,591, 5,876,930, 6,030,787, 6,258,569, and 5,804,375 (claims 1-12 only)). The amplicon sequence for each assay is however made available by ABI to reviewers on request.

DNA was amplified in 384 well microplates and in 4 µl reaction volumes containing 1 µl of 1 ng/µl DNA, 2 µl of 1x TaqMan Genotyping Master Mix (Applied Biosystems (ABI), Warrington UK), 0.2 µl of 20x assay mix (containing primers and probes from Applied Biosystems (ABI), Warrington UK) and 0.8 µl of sterile water. The thermal cycler conditions were: 10 minutes of pre-incubation at 95 ºC, followed by 40 cycles of 15 seconds at 92 ºC and 1 minute at 60 ºC. The resultant PCR product was analysed using TaqMan 7900HT software (Applied Biosystems (ABI), Warrington UK).

4.3 Results

4.3.1 Step 1: Diagnostic test to predict CYP1A2 functional variation among populations: test built from known *CYP1A2* variation

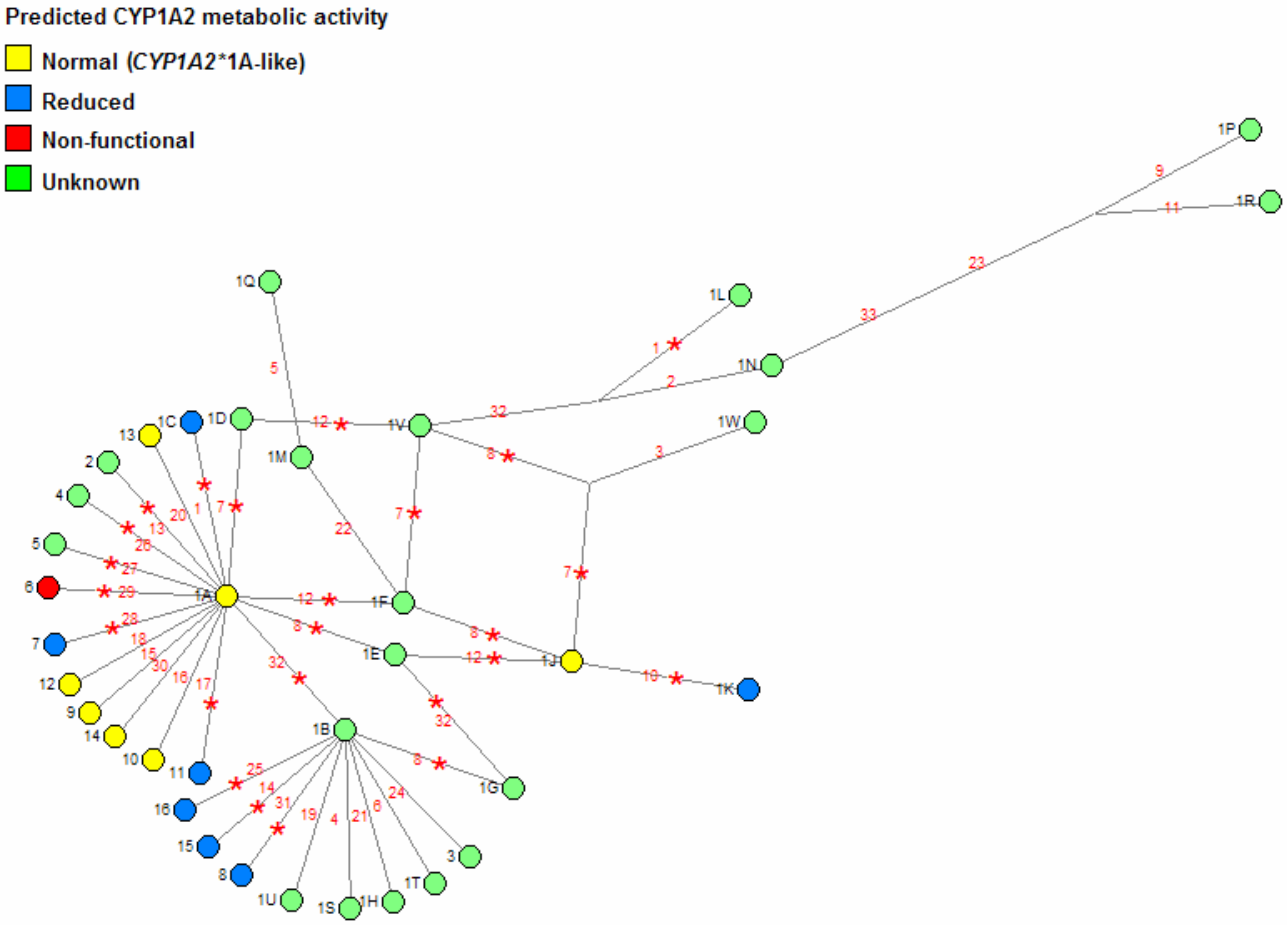
A mutation network was drawn using the *CYP1A2** alleles reported by the CYP450 Allele Nomenclature Committee (figure 4.1) in order to choose the minimum number of polymorphisms which need to be typed to distinguish between each of the independently evolved *CYP1A2* allele phenotypes. A total of 15 polymorphisms were selected (highlighted with * in figure 4.1).

The algorithm which will determine the predicted CYP1A2 phenotype based on the particular combination of alleles observed is shown in figure 4.2. As examples, chromosomes with -3860 G>A or -729 C>T are predicted to code for a slow functioning protein, whilst chromosomes with -2467 delT, or with -163 C>A but not -739 T>G and -729 C>T, are predicted to code for a protein whose function is not known.

Figure 4.2 Algorithm (based on known variation) for predicting CYP1A2 metabolic activity. White columns represent different chromosomes. √ = mutation must be present, x = mutation must be absent. Chromosomes which cannot be assigned to one of the predicted functions shown in the table are predicted to have a normal (*CYP1A2**1A-like) functioning *CYP1A2* allele.

Marker in figure 4.1	Mutation	Predicted CYP1A2 metabolic activity																
		Slow							Non-functional	Undetermined							Normal	
1	-3860G>A	√																
7	-2467delT												√					
8	-739T>G													x	√		√	
10	-729C>T		√											x	x		x	
12	-163C>A													√	x		√	
13	63C>G									√								
14	125C>G			√												x		
17	558C>A				√													
25	2473G>A					√										x		
26	2499A>T										√							
27	3496G>A											√						
28	3533G>A						√											
29	5090C>T								√									
31	5166G>A							√								x		
32	5347T>C															√		

Figure 4.1 Mutation network of *CYP1A2 alleles.** Nodes represent *CYP1A2** alleles reported by the CYP450 Allele Nomenclature Committee (http://www.cypalleles.ki.se/cyp1a2.htm (04/06/09). See table 2.6 (chapter 2) for references concerning allele phenotypes. Mutated positions are shown in red along the grey links. All links are drawn to scale. The network was used to help select the minimum number of polymorphisms (*) which need to be typed in order to predict CYP1A2 phenotypes. Reticulations are observed in the network with the consequence that some markers appear more than once.



Marker key

Marker	Mutation	Location	Amino acid change
*1	-3860G>A	5' region	
2	-3594T>G	5' region	
3	-3113A>G	5' region	
4	-3053A>G	5' region	
5	-2808A>C	5' region	
6	-2667T>G	5' region	
*7	-2467delT	5' region	
*8	-739T>G	Intron 1	
9	-733G>C	Intron 1	
*10	-729C>T	Intron 1	
11	-367C>T	Intron 1	
*12	-163C>A	Intron 1	
*13	63C>G	Exon 2	F21L
*14	125C>G	Exon 2	P42R
15	248C>T	Exon 2	T83M
16	502G>C	Exon 2	E168Q
*17	558C>A	Exon 2	F186L
18	634A>T	Exon 2	S212C
19	678C>T	Exon 2	F226F
20	1514G>A	Exon 3	G299S
21	1570A>C	Exon 3	A317A
22	2159G>A	Intron 4	
23	2321G>C	Intron 4	
24	2116G>A	Exon 4	D348N
*25	2473G>A	Exon 5	R377Q
*26	2499A>T	Exon 5	I386F
*27	3496G>A	Exon 6	C406Y
*28	3533G>A	Intron 6	
*29	5090C>T	Exon 7	R431W
30	5112C>T	Exon 7	T438I
*31	5166G>A	Exon 7	R456H
*32	5347T>C	Exon 7	N516N
33	5521A>G	3' UTR	

4.3.2 Step 2: CYP1A2 functional variation predicted from sequencing *CYP1A2* in the Ethiopian ascertainment populations

The enhancer region was not sequenced (see chapter 2) with the consequence that -3860 G>A and -2467 delT were not characterised via sequencing. Since -3860 G>A and -2467 delT were included in the algorithm and were needed to differentiate chromosomes into undetermined or reduced metabolic activity phenotypes, in this study, both SNPs were genotyped in the Ethiopian ascertainment populations. For each SNP, no population deviated significantly from Hardy Weinberg equilibrium at 5 % significance. Both SNPs were polymorphic in all populations (table 4.2).

Table 4.2 -3860 G>A and -2467 T>- frequencies in the Ethiopian ascertainment populations

	Afar		Amhara		Anuak		Maale		Oromo		Ethiopia	
SNP	n	f	n	f	n	f	n	f	n	f	n	f
-3860 G>A	26	0.19	21	0.14	43	0.29	24	0.17	28	0.21	142	0.20
-2467 T>-	7	0.05	6	0.04	30	0.21	11	0.08	12	0.08	66	0.09

n = number of chromosomes, f = frequency of the minor allele

A total of 60 *CYP1A2* haplotypes (including alleles at positions -3860 and -2467) were observed in the combined Ethiopian ascertainment population (table 4.3), 52 % of which were assigned a reduced CYP1A2 predicted metabolic activity, 3 % of which were predicted to code for a non-functional enzyme and 45 % of which were assigned an unknown CYP1A2 function. Haplotype distribution (table 4.4) varied among populations but haplotype 3 (*CYP1A2**1B), which was assigned an unknown function, was the most frequent haplotype in all groups (≥ 28 %).

4.3.3 Step 3: Application of the CYP1A2 diagnostic test (based on known data) to the Ethiopian ascertainment population, and the extent to which it predicts functional variation predicted from sequencing *CYP1A2*

Of the 15 *CYP1A2* SNPs included in the algorithm (from step 1), six were observed to be polymorphic in the Ethiopian ascertainment populations. These were: -3860 G>A, -2467 delT, -739 T>G, -729 C>T, -163 C>A and 5347 T>C.

Figure 4.3a shows the outcome of the diagnostic test once it was applied to the variation observed in the Ethiopian ascertainment populations in this study. Over 38 % of chromosomes in each population were predicted to code for a protein with a normal CYP1A2 function. Between 14 – 29 % of chromosomes in each group were predicted to code for proteins which function slowly but no non-functional alleles were predicted in any one group.

Figure 4.3b shows the CYP1A2 functional variation predicted from sequencing *CYP1A2* in the Ethiopian ascertainment populations. Similar to predictions from the diagnostic test (figure 4.3a)

Table 4.3 ***CYP1A2*** **haplotypes observed in the Ethiopian ascertainment populations and their predicted CYP1A2 metabolic activities** (frequencies are shown in table 4.4)

Nucleotide change [1]	-3860 G>A	-2467 T>-	-1014 C>A	-1008 G>A	-739 T>G	-729 C>T	-592 C>T	-505 G>A	-163 C>A	-151 G>T	53 C>G	1352 G>A	1370 G>A	1513 C>A	1589 G>T	1611 G>A	2159 G>A	2321 G>C	2534 C>T	3463 C>T	3468 A>C	3588 G>T	3605 A>G	3613 T>C	4957 C>G	4961 C>T	5010 C>T	5029 C>G	5094 T>C	5105 G>A	5253 C>G	5284 C>A	5328 G>A	5347 T>C	5355 G>C	5521 A>G	5620 A>C	5987 G>T	6233 C>T	6324 G>-	6388 T>G	6537 G>A	6562 T>A	6674 C>G	6685 A>G	6714 A>T	6765 C>T
Location	Enhancer region	Enhancer region	5' upstream	5' upstream	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Intron 1	Exon 2	Intron 2	Intron 2	Exon 3	Intron 3	Intron 3	Intron 4	Intron 4	Intron 5	Exon 6	Exon 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Intron 6	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	Exon 7	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR	3' UTR
Amino acid change											S18C			S298R						T395M	N397H								F432S	D436N	P485R	Y495Ter	R510Q	Synonymous													
Haplotype id [2] 1									■								■			■														■												
2									■								■																	■												
3																																		■												
4									■								■																	■												
5																					■													■												
6	■								■																									■										■		
7	■				■				■															■										■										■		
8																	■										■							■												
9	■	■							■																									■			■			■						
10	■								■									■																	■								■	■			
11																																	■	■												
12	■	■							■																									■												
13									■								■											■							■										■		
14	■								■																									■												
15	■				■				■															■										■	■	■								■		
16																																		■		■										
17									■																									■												
18												■																						■												
19	■								■									■																■		■								■		
20	■	■							■					■																				■				■			■					
21	■				■				■						■									■										■		■								■		
22									■																									■											■	
23	■				■				■																									■		■								■		
24	■	■	■	■					■								■																	■										■		
25	■	■							■	■																								■			■									
26	■				■				■							■								■										■		■								■		
27	■	■							■		■			■																				■												
28																														■				■											■	
29									■								■																	■												
30									■																■									■												
31									■								■													■				■												
32	■	■							■					■																				■												
33								■																										■												
34	■	■							■																									■										■		
35	■								■									■																■			■								■		
36																																		■										■		
37													■																						■												
38	■	■							■					■																			■												
39									■								■																			■												
40																																		■												
41	■	■							■																									■			■									
42	■	■							■					■					■														■												
43	■	■							■					■																				■										■		
44	■	■							■					■																				■				■								
45	■	■							■					■																				■				■								
46		■							■																									■												
47									■																									■												
48	■	■							■					■																				■			■									■
49									■																					■				■			■									
50	■				■				■						■									■										■		■			■					■		
51																							■						■					■										■		
52	■								■									■								■								■		■								■		
53	■	■							■					■																	■			■				■							
54	■				■	■			■															■										■		■								■		
55	■				■										■									■										■												
56									■														■											■		■								■		
57							■																						■					■												
58									■																									■												
59									■																									■												
60	■								■								■																	■												

[1] Position from base A in the initiation codon (A in ATG is +1, base prior to A is -1) from the *CYP1A2* genomic reference sequence (NC_000015.8)

■ **Reduced** ■ **Non-functional** ■ **Unknown**

[2] white cell = allele observed in *CYP1A2**1A, grey cell = derived allele

Predicted phenotypes: Haplotypes 38 and 42 were predicted to code for a truncated protein owing to the premature stop codon (P495Ter) in exon 7 (chapter 2). The -3860 G>A SNP in the enhancer region, which was reported as being causative of reduced CYP1A2 inducibility in Japanese smokers, most probably owing to reduced protein expression (Nakajima et al., 1999), was observed in just over half (52 %) of the haplotypes, which were assigned a reduced CYP1A2 metabolic activity as a result. Some of the haplotypes with -3860 G>A also included other variants associated with a reduced CYP1A2 metabolic activity: Haplotype 54 included -729 C>T which was reported to code for a reduced functional allele (Aklillu et al., 2003); haplotypes 27 and 53 included amino acid substitutions S18C (exon 2) and P485R (exon 7) respectively, both of which were predicted to have a damaging effect on the structure/function of the protein by PolyPhen software (chapter 2). Haplotype 31 was assigned a reduced CYP1A2 metabolic activity since it included the amino acid change F432S in exon 7 which was also predicted to have a damaging effect on the structure/function of the protein by PolyPhen software (chapter 2).

Table 4.4 *CYP1A2* haplotype distribution among the combined and individual Ethiopian ascertainment populations

Haplotype id	Afar		Amhara		Anuak		Maale		Oromo		Ethiopia	
	n	f	n	f	n	f	n	f	n	f	n	f
1	27	0.28	40	0.34	5	0.04	16	0.13	22	0.24	110	0.197
2	1	0.01	0	0.00	2	0.01	1	0.01	1	0.01	5	0.009
3	35	0.36	50	0.42	63	0.47	34	0.28	27	0.30	209	0.375
4	5	0.05	9	0.08	1	0.01	13	0.11	6	0.07	34	0.061
5	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	2	0.004
6	2	0.02	0	0.00	2	0.01	0	0.00	1	0.01	5	0.009
7	2	0.02	1	0.01	5	0.04	0	0.00	1	0.01	9	0.016
8	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	2	0.004
9	4	0.04	2	0.02	8	0.06	0	0.00	1	0.01	15	0.027
10	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	1	0.002
11	2	0.02	0	0.00	0	0.00	1	0.01	0	0.00	3	0.005
12	1	0.01	0	0.00	4	0.03	0	0.00	2	0.02	7	0.013
13	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	1	0.002
14	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	1	0.002
15	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	1	0.002
16	1	0.01	0	0.00	5	0.04	19	0.16	3	0.03	28	0.050
17	1	0.01	1	0.01	3	0.02	3	0.03	4	0.04	12	0.022
18	1	0.01	1	0.01	0	0.00	2	0.02	1	0.01	5	0.009
19	1	0.01	1	0.01	2	0.01	2	0.02	0	0.00	6	0.011
20	1	0.01	0	0.00	0	0.00	1	0.01	0	0.00	2	0.004
21	1	0.01	4	0.03	0	0.00	7	0.06	3	0.03	15	0.027
22	2	0.02	0	0.00	0	0.00	0	0.00	2	0.02	4	0.007
23	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	1	0.002
24	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	1	0.002
25	0	0.00	2	0.02	0	0.00	0	0.00	0	0.00	2	0.004
26	0	0.00	1	0.01	0	0.00	0	0.00	2	0.02	3	0.005
27	0	0.00	1	0.01	1	0.01	0	0.00	0	0.00	2	0.004
28	0	0.00	1	0.01	1	0.01	7	0.06	1	0.01	10	0.018
29	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	1	0.002
30	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	1	0.002
31	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00	1	0.002
32	0	0.00	0	0.00	8	0.06	1	0.01	2	0.02	11	0.020
33	0	0.00	0	0.00	10	0.07	0	0.00	0	0.00	10	0.018
34	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	1	0.002
35	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	1	0.002
36	0	0.00	0	0.00	3	0.02	0	0.00	0	0.00	3	0.005
37	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	1	0.002
38	0	0.00	0	0.00	3	0.02	0	0.00	0	0.00	3	0.005
39	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	1	0.002
40	0	0.00	0	0.00	1	0.01	1	0.01	0	0.00	2	0.004
41	0	0.00	0	0.00	1	0.01	1	0.01	0	0.00	2	0.004
42	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	1	0.002
43	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	1	0.002
44	0	0.00	0	0.00	0	0.00	4	0.03	1	0.01	5	0.009
45	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
46	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
47	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
48	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
49	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
50	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
51	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.002
52	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
53	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
54	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
55	0	0.00	0	0.00	0	0.00	0	0.00	2	0.02	2	0.004
56	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
57	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
58	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
59	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
60	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.002
Grand Total	96	1.00	118	1.00	134	1.00	120	1.00	90	1.00	558	1.000

n = number of chromosomes, f = frequency

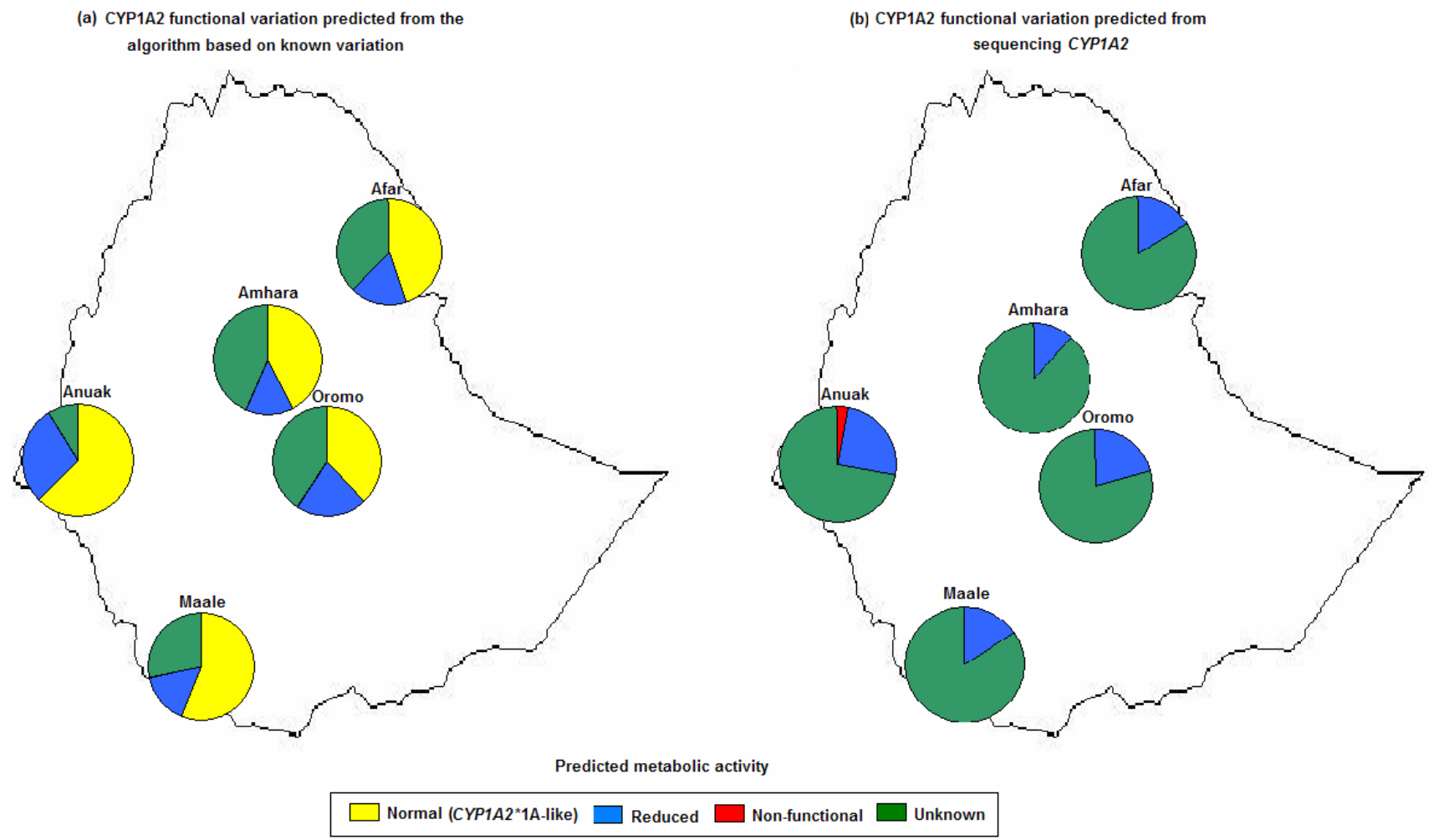
Predicted CYP1A2 metabolic activity:

Reduced

Non-functional

Unknown

Figure 4.3 Comparison of predicted CYP1A2 functional variation among the Ethiopian ascertainment populations. Pie charts represent the proportions of chromosomes in each population predicted to code for proteins with one of four CYP1A2 metabolic activities.



between 12 – 25 % of chromosomes in each population were predicted to code for a protein with reduced metabolic activity, with most (25 %) being observed in Anuak (figure 4.3b). Contrary to predictions from the diagnostic test however, normal functioning alleles were not predicted from *CYP1A2* sequences in any group, and 3 % of sequenced Anuak chromosomes were predicted to be non-functional (figure 4.3b). In addition, at least 72 % of sequenced chromosomes in each population were assigned an unknown CYP1A2 function.

Predictions from application of the diagnostic test (figure 4.3a) were consequently not consistent with the functional variation predicted from sequencing *CYP1A2* in the Ethiopian ascertainment populations (figure 4.3b). This is likely to be due to the plethora of previously unknown haplotypes identified from sequencing *CYP1A2* in the Ethiopian ascertainment populations (see chapter 2).

4.3.4 Step 4: Diagnostic test to predict CYP1A2 functional variation among Ethiopian populations: test built from Ethiopian sequence data from this study

A mutation network of the *CYP1A2* haplotypes observed in the Ethiopian ascertainment populations is shown in figure 4.4. Three polymorphisms need to be typed (-3860 G>A (1), 5094 T>C (29) and 5284 C>A (32)) in order to distinguish the alleles predicted to be poor functioning (highlighted in red) from the alleles which have an undetermined function (highlighted in green).

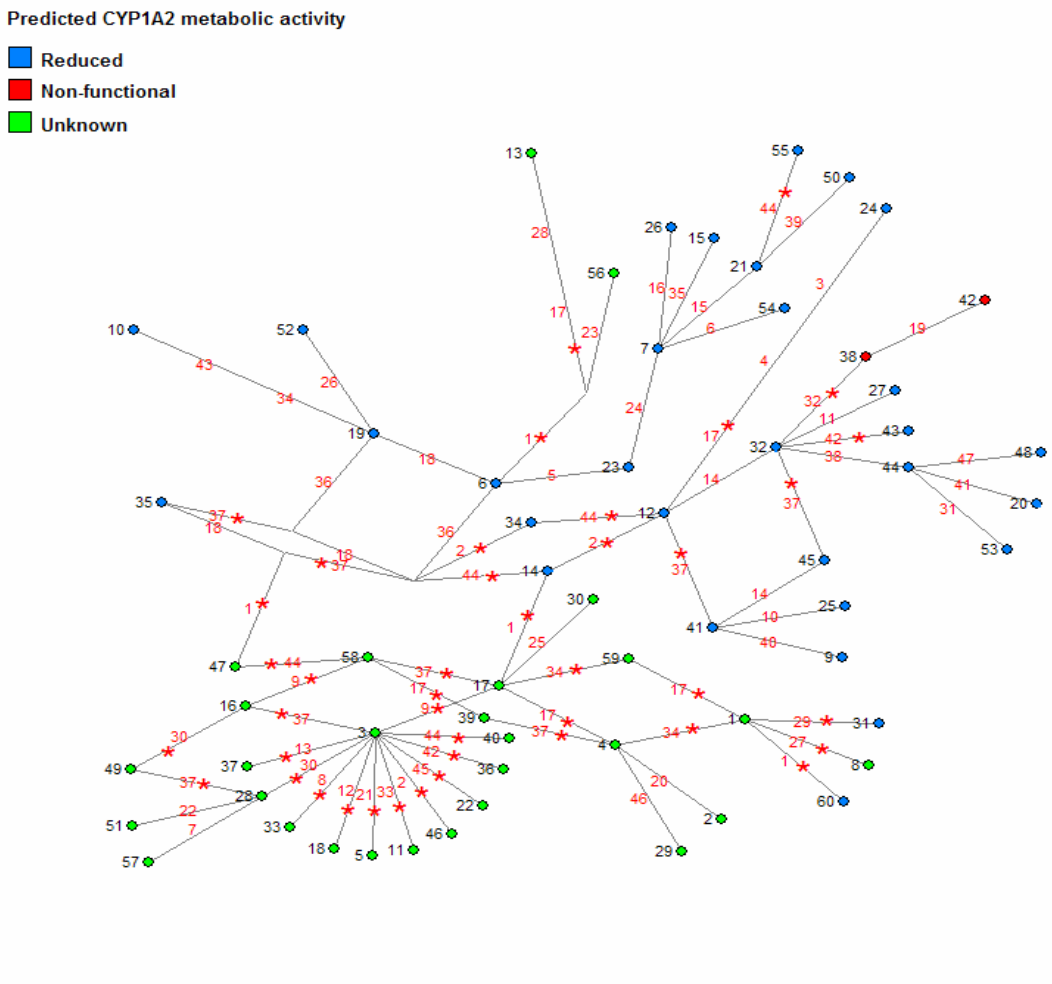
Since haplotypes 1 and 3 were observed at frequencies over 10 % in the combined Ethiopian ascertainment population (table 4.4) and were assigned an unknown function, these haplotypes should also be recognised by the diagnostic test for Ethiopia. Future genotype/phenotype studies should be performed using Ethiopian individuals with either haplotype. An additional 15 polymorphisms need to be typed to characterise haplotypes 1 and 3. These, together with -3860 G>A, 5094 T>C and 5284 C>A, are highlighted with * in figure 4.4.

The algorithm which will determine the predicted phenotype based on the particular combination of alleles observed is shown in figure 4.5. As examples, a chromosome with 5284 C>A is predicted to code for a protein with no function, whilst a chromosome with -3860 G>A but not 5284 C>A is predicted to code for a slow functioning protein.

4.4 Discussion

This study clearly established that it is not possible to predict functional CYP1A2 variation among Ethiopian populations using a diagnostic test built from variation recorded by the CYP450 Allele Nomenclature Committee (http://www.cypalleles.ki.se/cyp1a2.htm). Although the diagnostic test was derived from data from peoples of different ancestry (including Africans

Figure 4.4 Mutation network of *CYP1A2* alleles observed from sequencing *CYP1A2* in the Ethiopian ascertainment populations. See chapter 2 for references concerning allele phenotypes. Nodes represent *CYP1A2* alleles. Mutated positions are shown in red along the grey links. All links are drawn to scale. The network was used to help select the minimum number of polymorphisms (*) which need to be typed in order to predict CYP1A2 phenotypes. Reticulations are observed in the network with the consequence that some markers appear more than once.



Marker	Mutation	Location	Amino acid change
* 1	-3860 G>A	Enhancer region	
* 2	-2467 T>-	Enhancer region	
3	-1014 C>A	5' upstream	
4	-1008 G>A	5' upstream	
5	-739 T>G	Intron 1	
6	-729 C>T	Intron 1	
7	-592 C>T	Intron 1	
* 8	-505 G>A	Intron 1	
* 9	-163 C>A	Intron 1	
10	-151 G>T	Intron 1	
11	53 C>G	Exon 2	S18C
* 12	1352 G>A	Intron 2	
* 13	1370 G>A	Intron 2	
14	1513 C>A	Exon 3	S298R
15	1589 G>T	Intron 3	
16	1611 G>A	Intron 3	
* 17	2159 G>A	Intron 4	
18	2321 G>C	Intron 4	
19	2534 C>T	Intron 5	
20	3463 C>T	Exon 6	T395M
* 21	3468 A>C	Exon 6	N397H
22	3588 G>T	Intron 6	
23	3605 A>G	Intron 6	
24	3613 T>C	Intron 6	
25	4957 C>G	Intron 6	
26	4961 C>T	Intron 6	
* 27	5010 C>T	Intron 6	
28	5029 C>G	Intron 6	
* 29	5094 T>C	Exon 7	F432S
* 30	5105 G>A	Exon 7	D436N
31	5253 C>G	Exon 7	P485R
* 32	5284 C>A	Exon 7	Y495Ter
* 33	5328 G>A	Exon 7	R510Q
* 34	5347 T>C	Exon 7	Synonymous
35	5355 G>C	3' UTR	
36	5521 A>G	3' UTR	
* 37	5620 A>C	3' UTR	
38	5987 G>T	3' UTR	
39	6233 C>T	3' UTR	
40	6324 G>-	3' UTR	
41	6388 T>G	3' UTR	
* 42	6537 G>A	3' UTR	
43	6562 T>A	3' UTR	
* 44	6674 C>G	3' UTR	
* 45	6685 A>G	3' UTR	
46	6714 A>T	3' UTR	
47	6765 C>T	3' UTR	

Figure 4.5 Algorithm (based on *CYP1A2* sequence variation in the Ethiopian ascertainment populations) for predicting CYP1A2 metabolic activity. White columns represent different chromosomes. √ = mutation must be present, x = mutation must be absent. Chromosomes which cannot be assigned to one of the predicted functions shown in the table are predicted to have a *CYP1A2* allele whose associated function is not known.

Marker in figure 4.4	Mutation	Predicted CYP1A2 metabolic activity				
		Non-functional	Slow		Undetermined	
					Haplotype 1	Haplotype 3
1	-3860G>A		√		x	
2	-2467T>-					x
8	-505G>A					x
9	-163C>A					x
12	1352G>A					x
13	1370G>A					x
17	2159G>A				√	
21	3468A>C					x
27	5010C>T				x	
29	5094T>C			√	x	
30	5105G>A					x
32	5284C>A	√	x			
33	5328G>A					x
34	5347T>C				x	
37	5620A>C					x
42	6537G>A					x
44	6674C>G					x
45	6685A>G					x

and non-Africans), it was particularly biased towards the Japanese (Murayama et al., 2004; Soyama et al., 2005) and may consequently be suitable for predicting CYP1A2 functional variation in Japanese populations. The diagnostic test was found to be unsuitable for Ethiopia since it did not account for much of the diversity observed in the Ethiopian ascertainment populations. Ethiopia requires its own diagnostic test designed initially by using the variation observed in the five Ethiopian ascertainment populations.

Since it is not possible to confidently predict the likely phenotypes of many of the *CYP1A2* haplotypes identified in the Ethiopian ascertainment populations, functional studies will be necessary before a robust predictive algorithm can be formulated. The functional effect of the amino acid substitutions predicted to have a damaging effect on the structure/function of the protein should be established initially in *in vitro* studies. These are: 53C>G in exon 2 (S18C) and 5094T>C (F432S), 5253C>G (P485R) and 5284C>A (Y495Ter) in exon 7. Protein models could also be constructed *in silico* to establish whether SNPs alter protein shape/structure in any way (Karchin et al., 2005). Genotype/phenotype tests (using caffeine) investigating whether -3860 G>A (enhancer region) leads to reduced enzyme inducibility in Ethiopians should also be undertaken. Genotype/phenotype association studies should also be performed using haplotypes observed at appreciable frequencies. It may be appropriate, as a first milestone, to take into account haplotypes observed in the combined Ethiopian ascertainment population at frequencies ≥ 10 %. The functional significance of lower frequency haplotypes should be established in due course. It is also necessary to bear in mind that an overall frequency of less than 10 % might be significantly higher in one or more ethnic groups.

Haplotype 1 (*CYP1A2**1M) and haplotype 3 (*CYP1A2**1B) were the most frequent haplotypes in the combined Ethiopian ascertainment populations. Haplotype 1 has -163C>A in intron 1 and 2159G>A intron 4, whilst haplotype 3 has the synonymous mutation 5347T>C in exon 7 (N516N). Since these three mutations do not result in amino acid changes, both haplotypes may consequently code for a protein with a normal (*CYP1A2**1A-like) function. On the other hand -163C>A has been associated with higher enzyme inducibility by smoking (Sachse et al., 1999), although this association has not always been replicated (Nordmark et al., 2002; Aklillu et al., 2003), and haplotype 1 may code for an enzyme with an altered activity as a result. Genotype/phenotype association studies involving haplotypes 1 and 3 will establish whether functional variation exists because of variation within regulatory elements such as the promoter or enhancer region.

Future studies should conduct genotype/phenotype testing on approximately 100 individuals from the same ethnic group. A sample population of at least 100 individuals would be large enough to identify variants at frequencies of 2 % or above in a population at the 5 % significance threshold (refer to the discussion in chapter 2 for power calculation details). Ethiopians living in Ethiopia or the Ethiopian diaspora population in London could be involved in the study.

This study mainly focussed on variation within the gene since the enhancer region was not sequenced. Given that variation in the enhancer region may affect the quantity of active protein, future studies should also attempt to characterise its variation in the Ethiopian ascertainment populations and should incorporate this information into the design of the diagnostic test if it is to fully deliver its potential utility for group based pharmacogenetic prediction.

Table 4.5 Common African diseases, drugs used in their treatment in Ethiopia, and CYPs involved in the drugs' metabolism

Disease	Drug used to prevent/cure disease	Reference	Metabolised by	Reference
Malaria	Artemeter-lumefantrine	1, Ethiopia, MOH (2004)	CYP3A4, CYP2B6, CYP2D6	Giao and de Vries (2001); Li et al., 2003
Malaria	Chloroquine	1, Ethiopia, MOH (2004)	CYP3A4, CYP3A5, CYP2D6, CYP2C8	Kim et al., 2003; Li et al., 2003; Projean et al., 2003
Malaria	Primaquine	1, Ethiopia, MOH (2004)	**CYP1A2**, CYP2D6	Li et al., 2003
Malaria	Quinine	1, Ethiopia, MOH (2004)	CYP3A4, CYP2C19	Li et al., 2003, Mirghani et al., 2003
Malaria	Mefloquine	Ethiopia, MOH (2004)	CYP3A	Fontaine et al., 2000
Onchocerciaisis	Ivermectin	1	CYP3A4	Zeng et al., 1998
Schistosomiasis	Praziquantel	1	**CYP1A2**, CYP2C19, CYP2D6, CYP3A4, CYP3A5	Li et al., 2003

1 Information provided by Dr. Senbeta Guteta (medical faculty of Addis Ababa University) and Dr. Diriba Agegnehu (Benishangul-Gumz national regional state HIV secretariat Head)

Since many CYP enzymes have overlapping substrate specificities and enzymatic redundancy is evident in the CYP450 system (Gu et al., 1992), multiple CYPs would have to be included in the study in years to come, to fully predict the metabolic effect of the enzyme variants. As an

example, CYP1A2 and CYP2D6 enzymes both metabolise the anti-Malarial drug Primaquine (Li et al., 2003). As a consequence, the presence of a normally functioning CYP2D6 protein in an individual, may compensate for the absence or defective form of a CYP1A2 protein, and vice versa. Furthermore, true poor metaboliser phenotypes may arise from the presence of defective genes for both enzymes. Performing combined genotype analysis for defective *CYP1A2* and *CYP2D6* polymorphisms is thus important in determining the contribution of CYP genotypes to phenotype. Since CYP2B6, CYP2C8, CYP2C19, CYP2D6, CYP3A4 and CYP3A5 are thought to be involved in the metabolism of various anti-malarials and drugs used to treat Onchocerciasis and Schistosomiasis in Ethiopia (table 4.5), it would be appropriate to extend this study to these enzymes in the first instance.

4.5 Conclusion

This study examined the design of a diagnostic test procedure to predict CYP1A2 functional variation among Ethiopian populations. The test procedure was constructed utilising *CYP1A2* variant alleles, recorded by the CYP450 Allele Nomenclature Committee, which were assigned phenotypes from previously reported functional studies. The test procedure was derived from data from peoples of different ancestry, with a particular bias towards Japanese. The test's suitability for Ethiopia was investigated by applying it to the variation observed in the Ethiopian ascertainment populations. The diagnostic test was found to be inappropriate for Ethiopia since it did not account for the plethora of novel variation observed in the ascertainment samples. Ethiopia requires its own diagnostic test procedure initially built using the variation observed in the five Ethiopian ascertainment populations. The functional significance of all of the Ethiopian ascertainment haplotypes is currently not known and a robust diagnostic test procedure cannot be designed at present. Future studies should attempt to establish the function of potentially deleterious amino acid substitutions and high frequency haplotypes observed in the Ethiopian ascertainment populations. Future studies should also incorporate any variation in the enhancer region into the design of the diagnostic test procedure to improve its utility in group based pharmacogenetic prediction. The use of group based pharmacogenetic profiles for drug response prediction may be both a practical and beneficial approach in the pursuit of improved healthcare in Ethiopia and elsewhere.

5 The distribution of *TCF7L2* alleles, associated with an increased risk of type 2 diabetes, among Afro-Caribbeans with the disease, HapMap and Ethiopian populations

5.1 Introduction

5.1.1 Type 2 diabetes

Type 2 diabetes (T2D), otherwise known as diabetes mellitus ("sugar diabetes"), is a long-term condition characterised by hyperglycemia which is caused by impaired insulin secretion, insulin resistance in peripheral tissues and increased glucose output by the liver (Grant et al., 2006). The majority of type 2 diabetics suffer from chronic hyperglycemia which can lead to nephropathy, neuropathy, retinopathy and an increased risk of cardiovascular dieasease (Grant et al., 2006). T2D often occurs later in life and is known as late-onset diabetes or non-insulin-dependent diabetes mellitus (NIDDM), because insulin treatment is not always necessary and the disease can often be managed by exercise and diet (Burnet et al., 2006). T2D is a complex disease and many genes, in addition to environmental factors, are thought to determine an individual's risk to developing the disease (Weedon, 2007). The global prevalence of all types of diabetes for all age groups was estimated to be 2.8 % in 2000, with T2D being accountable for at least 90 % of this figure (Oldroyd et al., 2005). The prevalence of diabetes is set to substantially increase in the near future, largely because people are living longer (Wild et al., 2004) but also because of diet and lifestyle changes leading to obesity (Singh et al., 2004; Wiegand et al., 2004).

5.1.2 A genetic element to T2D

Despite having a strong environmental component, evidence from twin and family studies suggests that T2D risk also has an underlying genetic element (Gloyn and McCarthy, 2001). Furthermore, the prevalence of T2D varies widely amongst different ethnic groups, being particularly common among people of Afro-Caribbean or South Asian origin (Oldroyd et al., 2005). For example, by comparison to UK European whites with an estimated T2D prevalence of 2.4 %, the rate in those of Indian and Afro-Caribbean descent living in the UK is three- to sixfold higher (Riste et al., 2001). African Americans are also twice as likely to develop T2D compared to non-Hispanic white populations in the US (Harris et al., 1998).

5.1.3 The thrifty gene hypothesis

African and Asian diasporas are more likely to develop T2D than their respective source

populations, e.g. the prevalence of T2D is higher (~ 11-13 %) among people of recent African descent in developed countries than in Africa (~ 1-2 %) (Rotimi et al., 2004). It is however necessary to bear in mind that the possibility of T2D under diagnosis in developing countries cannot be ruled out. This phenomenon is suggested to be due to changes in environmental conditions acting on thrifty genes (genes promoting the efficient absorption, storage, or utilization of nutrients in periods of food abundance) (Neel, 1962). Prior to the development of farming roughly 10,000 years ago (Mithen, 2007), anatomically modern humans lived as hunter-gatherers and were likely to experience frequent cycles of food abundance and famine (Campbell and Tishkoff, 2008). According to the thrifty gene hypothesis (Neel, 1962), ancestral genetic variants that were historically advantageous in enabling individuals to efficiently process food to deposit fat during periods of food abundance, are now maladaptive to more modern environments with a constant abundance of food. The 'thrifty' genotype effectively prepares individuals for a famine that never comes, resulting in obesity and an increased risk of T2D (Di Rienzo and Hudson, 2005; Paradies et al., 2007).

5.1.4 Genes implicated in T2D aetiology

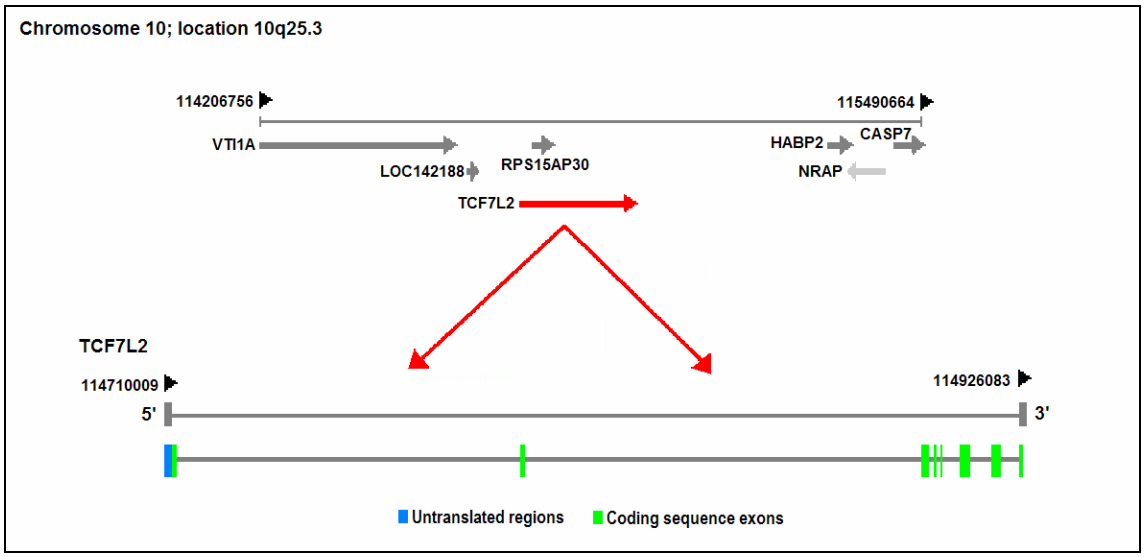
Recent genome-wide association studies in large sample sizes (tens of thousands of individuals) have identified several common variants (allele frequencies ranging from 10-90 %) associated with an increased susceptibility to T2D (Weedon, 2007). Variants include rs4402960 (*IGF2BP2*) and rs1801282 (*PPARG*) on chromosome 3, rs10946398 (*CDKAL1*) and rs10010131 (*WFS1*) on chromosome 6, rs13266634 (*SLC30A8*) on chromosome 8, rs10811661 (*CDKN2B*) and rs564398 (*CDKN2B*) on chromosome 9, rs5015480 (*HHEX*) on chromosome 10, rs5215 (*KCNJ11*) on chromosome 11, and rs8050136 (*FTO*) on chromosome 16 (Zeggini et al., 2007). With particular relevance to populations with a recent African ancestry, the porcine agouti-related protein (*AGRP*) (Bonilla et al., 2006) and proprotein convertase subtilisin/kexin-type 2 (*PCSK2*) (Leak et al., 2007) genes have been identified as candidate loci associated with diabetes. Genome-wide association studies in West Africans have also suggested association of regions of chromosomes 12, 19 and 20 to diabetes susceptibility (Rotimi et al., 2004). There have also been other reports of strong association with chromosomes 4, 6, 8, 10 and 15-18 for a number of traits that contribute to the disease (Chen et al., 2007; Chen et al., 2007; Rotimi et al., 2006).

5.1.4.1 The transcription factor 7-like 2 gene (*TCF7L2*)

TCF7L2, also known as *TCF-4*, is found on the long arm of chromosome 10 (figure 5.1). The gene is ~ 217.43 kb in length and has 10 transcripts which range from 7 exons (620 bp) to 16 exons (2838 bp) in length (Ensembl release 55, July 2009). *TCF7L2* is a high mobility group (HMG) box-containing transcription factor which plays a role in the Wnt signalling pathway, acting as a nuclear receptor for *CTNNB1* (β-catenin) (Smith, 2007). Wnt proteins form a family

of highly conserved secreted signalling molecules that are important in cell-to-cell interactions during cell proliferation, motility and normal embryogenesis (Etheridge et al., 2004). Wnt signalling has also been shown to regulate myogenesis and adipogenesis (Ross et al., 2000) and, with particular importance to glucose homeostasis, is critical for the development of pancreatic tissue and islets of Langerhans during embryonic development (Papadopoulou et al., 2005). Further support for the importance of Wnt signalling in glucose homeostasis stems from the recent finding that common variants within *HHEX* (target of Wnt signalling) also correlate with T2D risk (Weedon, 2007).

Figure 5.1 Location and structure of *TCF7L2*



5.1.4.2 Alleles of *TCF7L2* associate with an increased risk of T2D

In 2006, Decode Genetics demonstrated that the microsatellite marker, DG10S478, in exon 3 of *TCF7L2*, was significantly associated with T2D in Icelandic subjects (Grant et al., 2006). The association was also replicated in American and Danish cohorts. The strong link between DG10S478 and T2D risk was then replicated with other markers within *TCF7L2* including two intronic SNPs (rs7903146 C>T intron 5 and rs12255372 G>T in intron 6). The alleles were all strongly associated with each other and represented one risk allele present in ~ 36 % of individuals with T2D and ~ 28 % of control subjects. The increased risk of T2D was considerable, with each extra copy of the risk allele increasing the odds of the disease by 1.5-fold ($p \sim 10^{-18}$) (Grant et al., 2006).

Humphries et al., (2006) confirmed the association between rs7903146 and rs12255372 in European whites, Indians and Afro-Caribbeans, and found weaker associations for rs12255372 compared to rs7903146. LD between the two SNPs was high in Indians (D' = 0.94) and European whites (D' = 0.88), but considerably lower in Afro-Caribbeans (D' = 0.17). Haplotype frequencies also differed considerably in Afro-Caribbeans (e.g. both CT and TG were observed at frequencies of ~ 20 % in Afro-Caribbeans compared to < 6 % in European whites and

Indians. TT was the rarest haplotype in Afro-Caribbeans (9 %) yet observed at frequencies of ≥ 27 % in European whites and Indians). Only the TG haplotype was associated with an increased risk of T2D in Afro-Caribbeans (albeit not at 5 % significance, p = 0.24) whilst TG and TT were associated with a significantly increased risk of the disease in European whites and Indians. The study recommended further investigation into the TT and TG haplotypes.

The association of *TCF7L2* alleles with an increased risk of T2D has also been observed in other populations including Indians (Chandak et al., 2007), Japanese (Hayashi et al., 2007; Horikoshi et al., 2007), Mexican-Americans (Lehman et al., 2007), West Africans (Helgason et al., 2007), Moroccans (Cauchi et al., 2007), French (Sladek et al., 2007), Amish (Damcott et al., 2006) and Finnish (Scott et al., 2006). Of the known T2D candidate genes identified to date, *TCF7L2* makes the strongest contribution to T2D susceptibility and might consequently be the most important gene contributing to T2D risk (Weedon, 2007).

5.1.5 Aetiology of T2D

The causal variant of the association of *TCF7L2* and an increased risk of T2D is not known. Extensive sequencing of *TCF7L2* in Icelandic, Danish and American cohorts revealed no functional sequence change on the *TCF7L2* risk haplotype (Grant et al., 2006) and no functional candidates have been reported by other studies (Weedon, 2007). The SNP in intron 5 of the gene, rs7903146, which shows the strongest association (Humphries et al., 2006; Helgason et al., 2007), remains the most likely candidate. Since this variant occurs in an intron and has not been linked to any coding variation to date (Weedon, 2007), the causal variant is likely to act by altering protein expression, as opposed to protein structure. The causal variant's effect on T2D risk may act through an impairment of insulin secretion (Weedon, 2007). Individuals with rs7903146 T had significantly reduced levels of insulin secretion compared to the ancestral C allele (Florez et al., 2006) and TT homozygotes had ~ 50 % lower insulinogenic and insulin deposition indexes (Saxena et al., 2006). Further support for a role of *TCF7L2* in insulin regulation stems from a birth weight study. Correcting for foetal genotype, each maternal copy of rs7903146 T allele was associated with a 31 g increase in offspring birth weight (Freathy et al., 2007). The increase in offspring birth weight is thought to have resulted from the rs7903146 T allele reducing maternal insulin secretion, thereby elevating maternal blood glucose levels during pregnancy (Freathy et al., 2007).

The mechanism by which a reduced or non-functional allele of *TCF7L2* should impair insulin secretion remains unclear. Given that *TCF7L2* is part of the Wnt signalling which is critical for the development of the pancreas and islets during embryonic growth, impairment of beta-cell mass, pancreatic beta-cell development and/or beta-cell function is possible (Weedon, 2007). However, since *TCF7L2* has been linked to the development of colorectal cancer (Wong and Pignatelli, 2002), and *TCF7L2* homozygous knockout mice lack an epithelial stem-cell compartment, and die ~ 24 hours after birth (Korinek et al., 1998), it has been hypothesised that

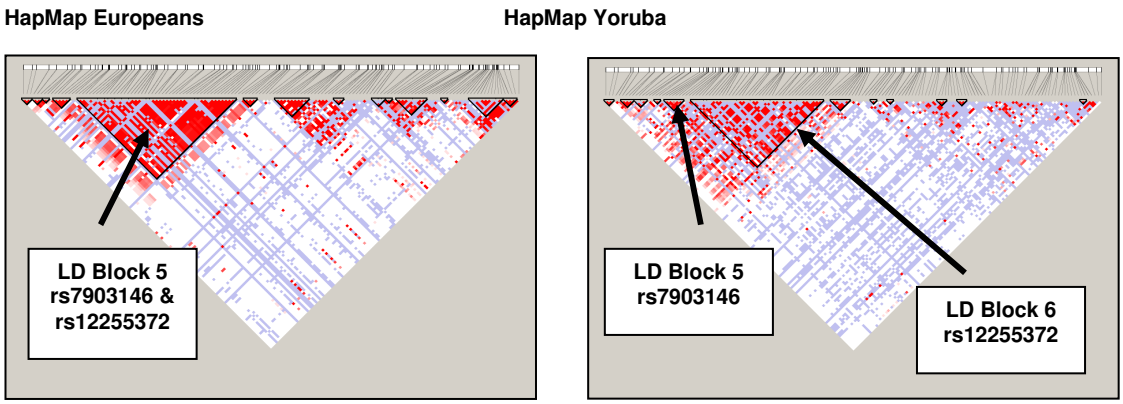
the defect may result from a decrease in glucagon-likepeptide-1 secretion from enteroendocrine cells (Grant et al., 2006). (GLP1 is an insulinotropic hormone which is transcriptionally regulated by *TCF7L2* (Yi et al., 2005)).

Identifying the causal genetic variant of the association of *TCF7L2* with an increased risk of T2D, and being able to specifically target treatments to carriers of this defect, would lead to improved management and treatment of people with or at risk of developing this complex disease (Weedon, 2007).

5.1.6 *TCF7L2* in HapMap European and Yoruba populations

LD was analysed across the entire *TCF7L2* gene in HapMap European and Yoruba populations using Haploview (figure 5.2). SNPs rs7903146 and rs12255372 were found in the same LD block in Europeans, where the LD between the two SNPs was high (D´ = 0.95). The SNPs were found in separate LD blocks in Yoruba; rs7903146 was located in LD block 5 and rs12255372 in LD block 6. This was not surprising given the low LD value between the two markers (D´ = 0.14).

Figure 5.2 LD across *TCF7L2* in HapMap populations (red = high LD, white = low LD)



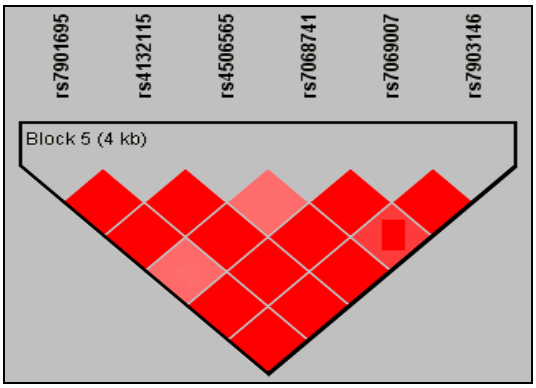
5.1.7 Aims

1. Analyse variation in *TCF7L2* in an Afro-Caribbean cohort with T2D

- Refine the search for a potentially causal polymorphism of T2D in Afro-Caribbeans from 55053 bases (distance between rs7903146 and rs12255372) to a smaller area in *TCF7L2*. Since Afro-Caribbeans may have a predominantly West African ancestry (Miljovic-Gacic et al., 2005), use HapMap West African Yoruba data to help narrow the search. Given that rs7903146 and rs12255372 are in separate LD blocks in HapMap Yoruba (figure 5.2), and rs7903146 C>T was more strongly associated with an increased risk of T2D than

rs12255372 G>T (Humphries et al., 2006), limit the search to LD block 5 of HapMap Yoruba (figure 5.3).

Figure 5.3 LD block 5 of *TCF7L2* in HapMap Yoruba (red = high LD)



- Analyse the distribution of haplotypes (consisting of SNPs in LD block 5 of the Yoruba (figure 5.3) and rs12255372) in a type 2 diabetic cohort from Humphries et al. (2006). Using HapMap data (Yoruba, Europeans, Chinese and Japanese) as a control, establish whether a potential T2D risk haplotype, on the background of the TG rs7903146/rs12255372 haplotype, can be identified in Afro-Caribbeans with T2D.

2. Analyse variation in *TCF7L2* in the Ethiopian ascertainment populations

- Type rs7901695, rs4132115, rs4506565, rs7068741 and rs7069007, in addition to rs7903146 and rs12255372 in the Ethiopian ascertainment populations. Estimate allele and haplotype frequencies in each population.

- Calculate LD between SNPs in each population.

- Characterise variation within and among populations, and in the context of Afro-Caribbeans with T2D and HapMap data (from aim 1).

- Establish whether variation in *TCF7L2* can be used to effectively differentiate ethnic groups.

3. Analyse variation in *TCF7L2* in multiple Ethiopian populations

In the course of this study it was established that rs7903146 and rs12255372 haplotypes could effectively differentiate populations (see results). As a consequence, both SNPs were typed in multiple Ethiopian populations. The aims of this section are to:

- Estimate allele and haplotype frequencies in each population.

- Calculate LD between SNPs in each population.

- Characterise variation within and among populations, and in the context of Afro-Caribbeans with T2D and HapMap data (from aim 1).

5.2 Methods

5.2.1 Samples

DNA samples (n = 81) from Afro-Caribbeans with T2D are described in Humphries et al., (2006) whilst DNA samples from the Ethiopian ascertainment populations (Afar, Amhara, Anuak, Maale and Oromo) are described in chapter 2. Samples from other Ethiopian populations (including the Ethiopian ascertainment samples plus more individuals from Afar, Amhara, Anuak, Maale and Oromo) were collected according to the methods outlined in chapter 2 for the Ethiopian ascertainment populations.

5.2.2 Multiple Ethiopian populations

A total of 6350 Ethiopian individuals were genotyped for rs7903146 and rs12255372. This dataset included samples from multiple Ethiopian populations collected from various locations in a rough south west to north east transect across Ethiopia (figure 5.4). Whilst all samples were analysed in terms of Ethiopia overall, at the ethnic group level, only populations with at least 15 individuals were included in the analyses (47 ethnic groups analysed).

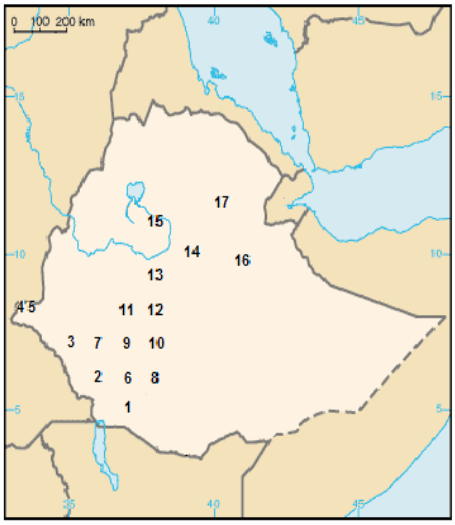
5.2.2.1 Marginalised groups study

Manja (n = 344) and Dawuro (n = 634), and Ari cultivators (n = 464), Ari potters (n = 262) and Ari smiths (n = 106) form part of a wider study which researches the genetic relationships of groups living in hierarchical societies (groups highlighted with * in figure 5.4). In this particular case, Manja and Dawuro are different ethnic groups living side by side in the Dawuro zone. Dawuro (farmers) are considered to be the dominant social group whilst Manja are the subordinate ethnic minority involved with artisanal production and hunting. Manja are heavily discriminated against and it is custom for the two groups not to interact with each other (Freeman and Pankhurst, 2003). Ari cultivators, Ari potters and Ari smiths are all from the Ari ethnic group living near Jinka town in South Omo. They are separate from the Ari dataset (n = 232) which does not consider social structure evident in Ari. Ari cultivators (farmers) are the dominant social group whilst, as a consequence of their trade, Ari potters (potters) and Ari smiths (blacksmiths) are the subordinate ethnic minorities. The potters and smiths are marginalised groups who, despite making important economic and social contributions to the local community, are subjected to various forms of discrimination. As with Manja and Dawuro, it

Figure 5.4 Multiple Ethiopian populations: sample size (chromosomes), collection location and first languages

Population	n	Collection location	1st languages spoken in population (%)
Gewada	232	1	99% GW, 1% AM
Dizi	264	2	88% DI, 12% AM
Mejenger	232	3	100% MG
Nuer	238	4	100% NR
Anuak	218	5	100% AN
Ari	232	6	87% AA, 13% AM
Ari Cultivator *	464	6	96% AA, 4% AM
Ari Potter *	262	6	100% AA
Ari Smith *	106	6	100% AA
Bena	30	6	87% BE, 13% AM
Basketo/Masketo	228	6	96% BK, 3% AM, 1% AA
Bume/Nyangagtom	48	6	96% BM, 4% AM
Busa	254	6	84% BU, 6% GM, 3% ZS, 2% DR, 2% GB, 1% KS
Dime	82	6	93% DM, 7% AM
Dirasha	216	6	65% DR, 26% AM, 8% OR, 1% ZS
Dasenech	54	6	89% DS, 11% AM
Gobeze	238	6	80% GB, 14% OR, 5% MH, 1% DR
Gofa	226	6	57% AM, 34% GF, 8% GM, 1% HM, 1% TG
Ganjule	220	6	100% GG
Gamo	410	6	66% AM, 33% GM, 1% WL
Genta	226	6	71% GN, 18% AM, 7% GM, 2% OR, 2% GF
Hamer	46	6	78% HM, 22% AM
Konso	242	6	76% KS, 20% AM, 2% OR, 1% GM, 1% WL, 1% ZS
Mashile	278	6	87% MH, 12% OR, 1% GB
Maale	244	6	93% ML, 4% AM, 2% GF, 1% AA
Wolayta	224	6	58% WL, 32% AM, 8% GM, 1% GR-SD, 1% OR
Zayse	222	6	89% ZS, 10% AM, 1% GM
Bench	256	7	98% BN, 1% AM, 1% KF
Kefa	242	7	89% KF, 8% AM, 2% OR
Shekecho	248	7	94% SC, 4% AM, 2% OR
Sheko	226	7	99% SK, 1% AM
Burji	242	8	72% BR, 26% OR, 1% KR, 1% KS
Dorze	208	8	95% DZ, 2% AM, 1% GM, 1% GZ, 1% OR
Gedeo	246	8	99% GE, 1% OR
Kore	228	8	68% KR, 17% OR, 14% AM, 1% GM
Dawuro *	634	9	97% DW, 2% AM, 0.5% GM, 0.5% WL
Konta	210	9	100% KN
Manja *	344	9	99% MA, 1% DW
Alaba	232	10	93% AL, 6% AM, 1% GR-SL
Kembata	232	10	90% KM, 9% AM, 1% OR, 1% WL
Sidama	254	10	87% SD, 13% AM
Oromo	306	11	73% OR, 27% AM
Gurage	308	12	52% AM, 18% GR-MS, 14% GR-SD, 10% GR-SL, 4% GR, 1% GR-KN, 1% GR-SB, 1% MR-HD, 1% OR
Hadiya	260	12	82% HD, 15% AM, 2% MR-HD, 1% GR-SL
Yem	218	12	99% YM, 1% AM
Tigrayan	140	13	50% AM, 49% TG, 1% AG
Amhara	822	14	100% AM
Agew, Eastern	542	15	56% AG, 44% AM
Somali	218	16	92% SM, 6% AM, 1% HR, 1% OR
Afar	232	17	99% AF, 1% AM

Ethiopian collection locations key



Language key

Language code	Language	Linguistic group
AA	Ari/Arigna	Omotic
AE	Arbore/Arborigna	Cushitic
AF	Afar/Afarigna	Cushitic
AG	Agew/Agewigna	Cushitic
AL	Alaba/Alabigna	Cushitic
AM	Amhara/Amharic	Semitic
AN	Anuak/Anuakigna	Nilo-Saharan
BE	Bena/Benigna	Omotic
BK	Basketo/Basketigna	Omotic
BM	Bume/Bumigna	Nilo-Saharan
BN	Bench /Benchigna	Omotic
BR	Burji/Burjigna	Cushitic
BU	Busa/Busigna	Cushitic
DI	Dizi/Dizi	Omotic
DM	Dime/Dimegna	Omotic
DR	Dirasha/Dirashgna	Cushitic
DS	Dasenech/Dasenechigna	Cushitic
DW	Dawuro/Dawrogna	Omotic
DZ	Dorze/Dorzigna	Omotic
GB	Gobeze/Gobezegna	Cushitic
GE	Gedeo/Gedeogna	Cushitic
GF	Gofa/Gofigna	Omotic
GG	Ganjule/Ganjuligna	Omotic
GL	Ghleb/Ghlebigna	Cushitic
GM	Gamo /Gamogna	Omotic
GN	Genta/Gentigna	Omotic
GR	Gurage/Guragigna	Semitic
GR-KN	Gurage-Kistane	Semitic
GR-MS	Gurage-Meskan	Semitic
GR-SB	Gurage-Sebat Bet	Semitic
GR-SD	Gurage-Sodo	Semitic
GR-SL	Gurage-Silte	Semitic
GW	Gewada/Gewadigna	Cushitic
GZ	Geez/Geez	Semitic
HD	Hadiya/Hadiyigna	Cushitic
HM	Hamer/Hamerigna	Omotic
HR	Harari (Adere)	Semitic
KF	Kefa/Kefigna	Omotic
KM	Kembata/Kembatigna	Cushitic
KN	Kistane/Kistanigna (Gurage)	Semitic
KO	Karo/Karogna	Omotic
KR	Kore/Korigna	Omotic
KS	Konso /Komsigna	Cushitic
MA	Manja/Manjigna (Keligna)	Omotic
MG	Mejenger/Mejengerigna	Nilo-Saharan
MH	Mashile (Mashole)	Cushitic
MJ	Maji	Omotic
MK	Mesketo/Mesketigna	Omotic
ML	Malle/maalegna	Omotic
MR-HD	Marako-Hadiya	Cushitic
MU	Mursi/Mursigna	Nilo-Saharan
NR	Nuer/Nuerigna	Nilo-Saharan
OR	Oromo/Oromigna	Cushitic
SC	Shekecho/Shekigna	Omotic
SD	Sidama/Sidamina	Cushitic
SK	Sheko/Shekogna	Omotic
SM	Somali/Somaligna	Cushitic
TG	Tigrian/Tigrigna	Semitic
TY	Tsemai/Tsemaina	Cushitic
WL	Wolayta/Wolaytigna	Omotic
YM	Yem (i)/Yemigna	Omotic
ZS	Zayse/Zaysigna	Omotic

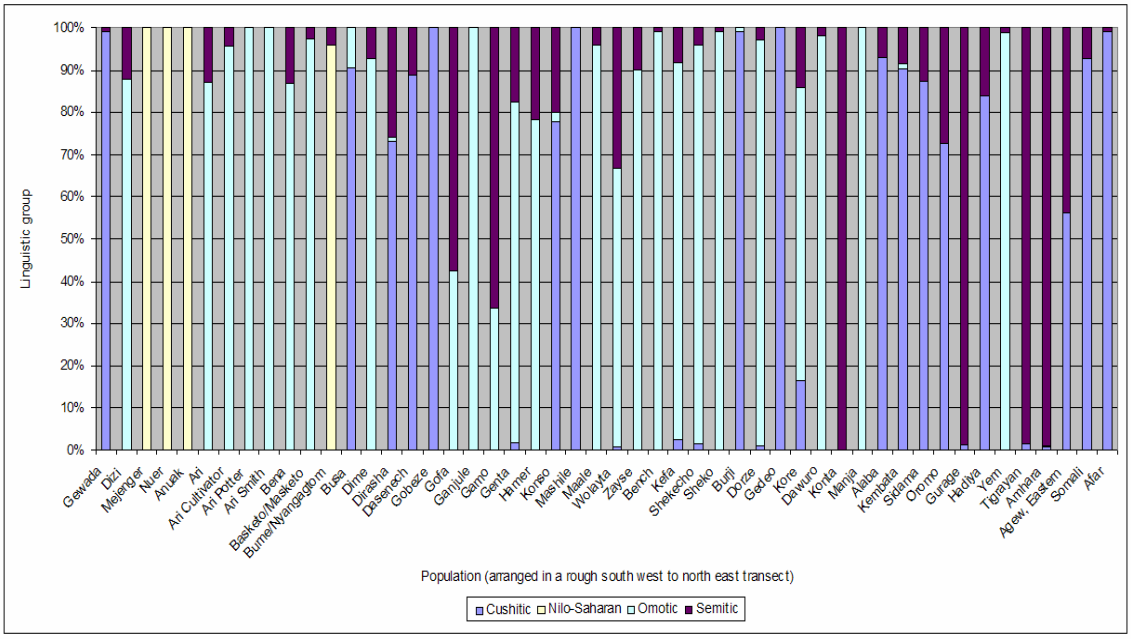
* Populations included in the marginalised groups study

is also customary for Ari farmers not to interact with Ari potters and smiths (Freeman and Pankhurst, 2003).

5.2.2.2 Language

In total, 62 first languages (mother tongue), from four linguistic groups, were reported when all ethnic groups were pooled (figure 5.4). The majority (40 %) of languages belonged to the Omotic family, whilst 34 % belonged to the Cushitic family and 18 % and 8 % belonged to the Semitic and Nilo-Saharan families respectively. Each ethnic group spoke their own language with most individuals in each group having their group's language as their mother tongue. For the majority of groups, several first languages were reported with many speaking Amharic. Semitic, Omotic and Nilo-Saharan languages were generally spoken more towards the north east, south west and west of Ethiopia respectively whilst Cushitic languages were generally spoken more in the south west and north east of the country (figure 5.5).

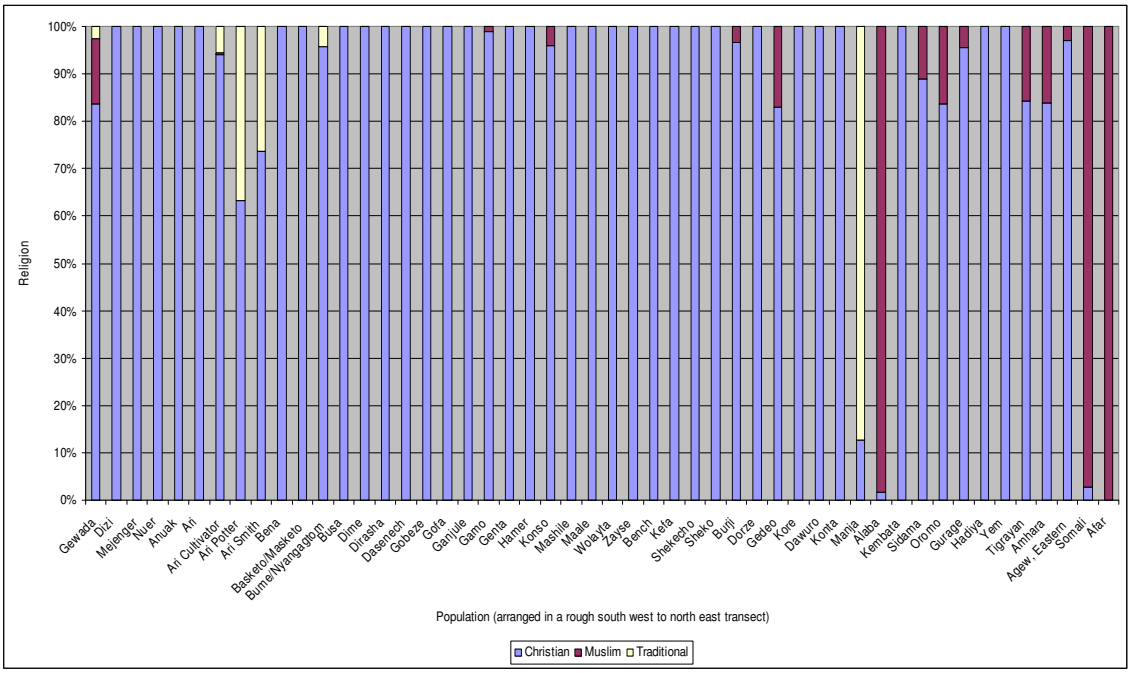
Figure 5.5 Distribution of first spoken languages (categorized by linguistic group) among multiple Ethiopian populations



5.2.2.3 Religion

In terms of the pooled Ethiopian dataset, the majority (88 %) were Christian, whilst 8 % were Muslim and 4 % had traditional religious beliefs. Christianity was observed in all ethnic groups except Afar who were all Muslim (figure 5.6). Islam was reported to be more widespread among groups towards the north east of Ethiopia and Traditional religions were observed more widely in marginalised groups (Manja, Ari potters and Ari smiths).

Figure 5.6 Distribution of religions among multiple Ethiopian populations



5.2.3 Genotyping of *TCF7L2* SNPs

Genotyping for all *TCF7L2* SNPs was performed using TaqMan technology (Applied Biosystems (ABI), Warrington UK). Forward and reverse primers, and fluorogenic probes are provided in table 5.1. For each assay, the design of the probes was such that if they annealed specifically between the forward and reverse primers, a sequence specific signal was generated as a direct result of probe degradation during PCR by the 5' nuclease activity of the polymerase.

DNA was amplified in 384 well microplates and in 4 µl reaction volumes containing 1 µl of 1 ng/µl DNA, 2 µl of 1x TaqMan Genotyping Master Mix (Applied Biosystems (ABI), Warrington UK), 0.01 µl of 40x assay mix (containing primers and probes from Applied Biosystems (ABI), Warrington UK) and 0.90 µl of sterile water. The thermal cycler conditions were: 10 minutes of pre-incubation at 95 ºC, followed by 40 cycles of 15 seconds at 92 ºC and 1 minute at 60 ºC. The resultant PCR product was analysed using TaqMan 7900HT software (Applied Biosystems (ABI), Warrington UK).

5.2.4 Statistical analyses

Haplotypes were inferred from unphased genotype data using Phase 2.0 software (Stephens et al., 2001). Mantel tests and partial mantel tests were performed using vegan libraries within the R statistical package (www.R-project.org). Population pairwise geographic distances were calculated based on great-circle differences. Population pairwise language, linguistic group and religious distances (Fsts) (Reynolds et al., 1983) were calculated from counts of languages,

Table 5.1 TaqMan primers and probes for *TCF7L2* genotyping

SNP	SNP alleles	Location in *TCF7L2*	Assay designed on forward or reverse strand	Forward primer	Reverse primer	VIC labelled probe	FAM labelled probe
rs7901695	C>T	Intron 5	Reverse	TGGATTTGCCTGTTCTTGACATTTC	TGAGAACCGTATGCTAAGTGAAAGC	AAAAGCCCGTAGATTT	CAAAAGCCCATAGATTT
rs4132115	C>A	Intron 5	Reverse	AGGCGCCAGCAAGCA	AGGCCCTCTGGGAAACCT	CCTTTCCTGATAGTCAC	CTTTCCTGAGAGTCAC
rs4506565	T>A	Intron 5	Reverse	TGACTCTCGGAGGAGGATGAG	GATGCCATAATAGAGACCCTTGACAA	CCCCATATCACTTCGG	CCCAAATCACTTCGG
rs7068741	C>T	Intron 5	Forward	ACCCCAGGATCCAGATGGTT	AGCTTATGCTATTGGTTCCAACACT	ACTCTCAAACCAGGGCC	ACTCTCAAATCAGGGCC
rs7069007	G>C	Intron 5	Forward	GATCCAGATGGTTGCACTCTCAA	GTCCTGACCAAAGCTTATGCTATTG	CCTTCAGCATTTAGTG	CCTTCACCATTTAGTG
rs7903146	C>T	Intron 5	Forward	CCTCAAAACCTAGCACAGCTGTTAT	TGAAAACTAAGGGTGCCTCATACG	CTAAGCACTTTTTAGATACTATAT	TAAGCACTTTTTAGATATTATAT
rs12255372	G>T	Intron 6	Forward	GCTGAGCTGCCCAGGAATAT	GCAGAGGCCTGAGTAATTATCAGAA	CAGGCAAGAATGACCATA	CCAGGCAAGAATTACCATA

linguistic groups and religions, respectively, in Arlequin software 3.01 (Schneider et al., 2000). In this case, languages, linguistic groups and religions were treated as if they were haplotypes. All other statistical analyses were performed according to methods outlined in chapter 2.

5.3 Results

5.3.1 Variation in *TCF7L2* in Afro-Caribbeans with T2D

In order to establish whether a potential T2D risk haplotype could be identified on the background of the TG rs7903146/rs12255372 haplotype in Afro-Caribbeans with T2D, SNPs within LD block 5 of HapMap Yoruba (rs7901695, rs4132115, rs4506565, rs7068741 and rs7069007 and rs7903146) together with rs12255372, were analysed. HapMap Yoruba, European, Chinese and Japanese populations were used as controls, and were assumed to be representative of their respective general populations (some may be susceptible to T2D, some may not). If this region of *TCF7L2* is associated with an increased risk of T2D, expectations were that a different distribution of SNP alleles or haplotypes would be observed in Afro-Caribbeans with the disease compared to the HapMap populations.

5.3.1.1 *TCF7L2* allele frequencies in Afro-Caribbeans with T2D and HapMap populations

For each SNP, no population deviated significantly from Hardy Weinberg equilibrium at 5 % significance. All seven SNPs were common (≥ 15 %) in Afro-Caribbeans with T2D, and were observed at frequencies most similar to Yoruba (figure 5.7).

5.3.1.2 *TCF7L2* haplotypes in Afro-Caribbeans with T2D and HapMap populations

Of the HapMap populations, Yoruba were the most similar to Afro-Caribbeans with T2D in terms of their haplotype distribution (figure 5.8a). All populations were significantly different (at 5 % significance) in terms of their haplotype frequencies, except for Afro-Caribbeans and Yoruba, and Chinese and Japanese (figure 5.8c). Furthermore, no haplotype appeared to be over-represented in Afro-Caribbeans compared to Yoruba. Only two Afro-Caribbean haplotypes (TCATGCG and CATCGTG) were not observed in Yoruba and both of these haplotypes were only observed in Afro-Caribbeans at a frequency of 1 % (figure 5.8a). The distribution of haplotypes on the background of rs7903146 T/rs12255372 G (named TG at risk allele hereafter) was also similar between Afro-Caribbeans with T2D and Yoruba (figures 5.8b and 5.8d).

Figure 5.7 ***TCF7L2*** **SNP allele frequencies in Afro-Caribbeans with T2D and HapMap populations** (derived alleles are always shown in purple)

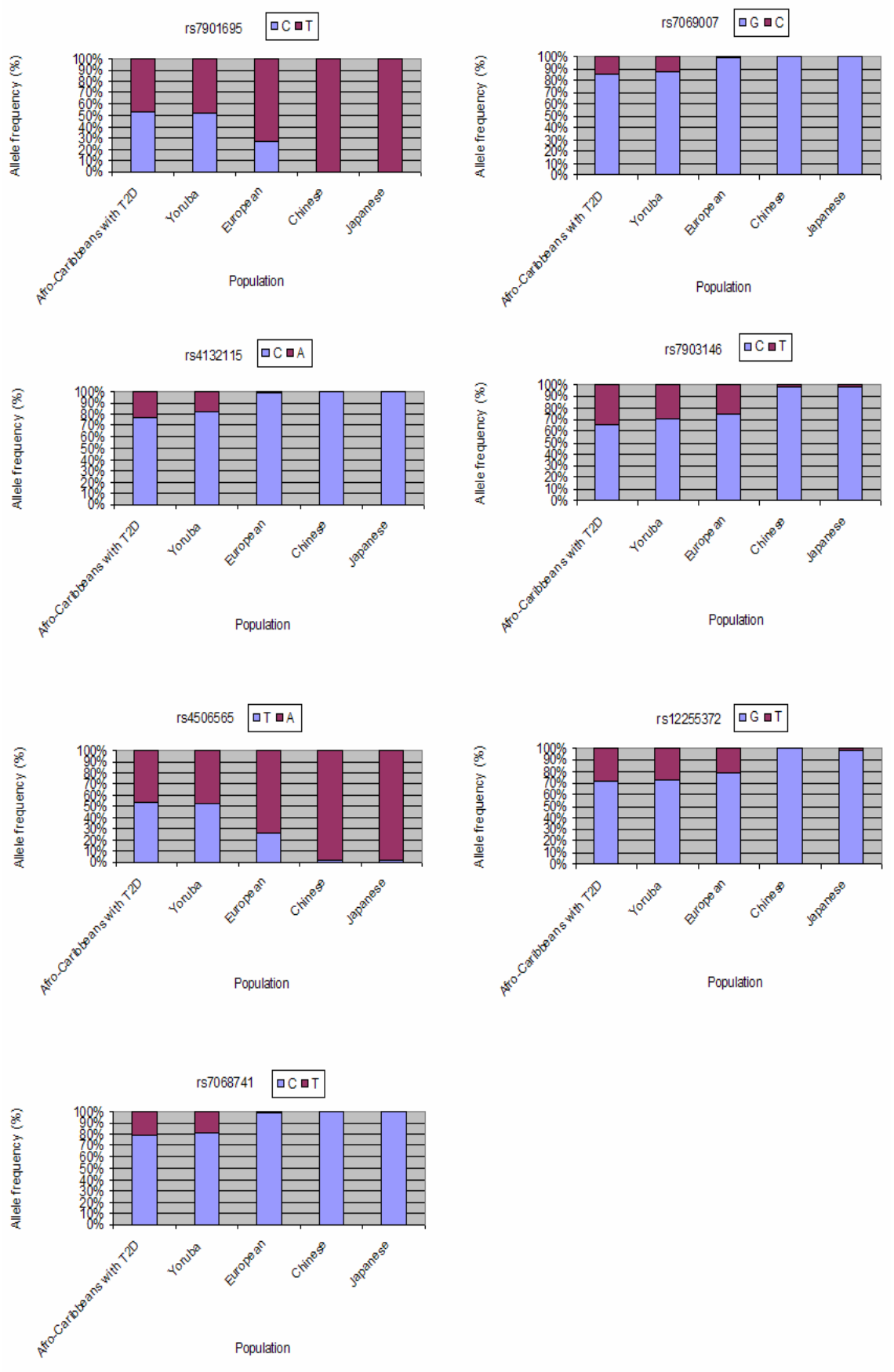
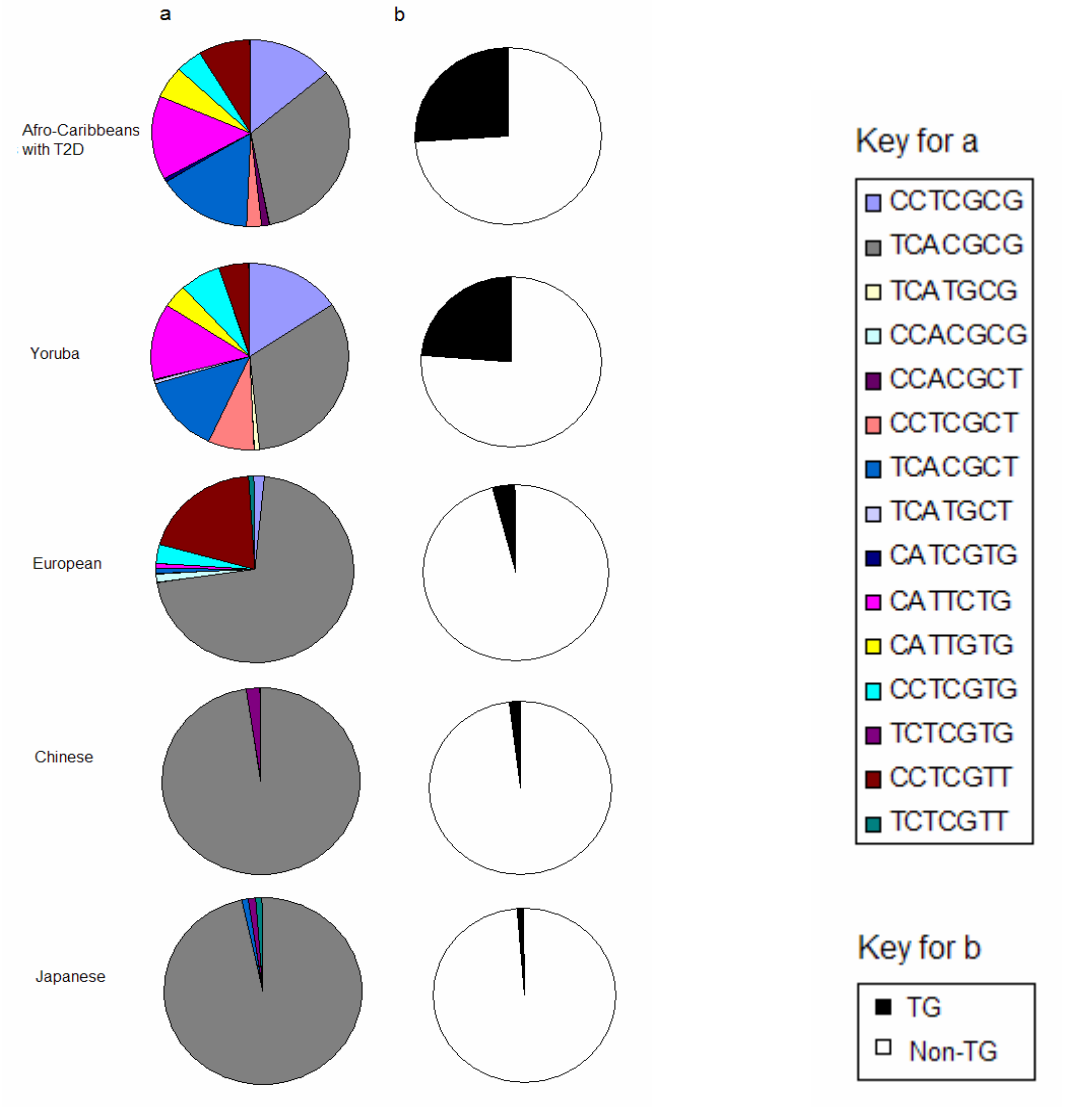


Figure 5.8 Comparison of *TCF7L2* haplotypes among Afro-Caribbeans with T2D and each of the HapMap populations. Frequencies of haplotypes (rs7901695, rs4132115, rs4506565, rs7068741, rs7069007, rs7903146 and rs12255372) are shown in figure 5.8a. Proportions of haplotypes ending with rs7903146/rs12255372 TG (those associated with an increased risk of T2D in Afro-Caribbeans (Humphries et al., 2006)) are shown in figure 5.8b.



c Exact test of population differentiation significant/not significant (+/-) differences at the 5 % threshold for *TCF7L2* haplotypes shown above (a)

	Afro-Caribbeans with T2D	Yoruba	European	Chinese	Japanese
Afro-Caribbeans with T2D					
Yoruba	-				
European	+	+			
Chinese	+	+	+		
Japanese	+	+	+	-	

d Pairwise Fisher's exact tests comparing numbers of haplotypes 'ending with TG' versus those 'not ending with TG' shown above in (b) **- = p > 0.05, +++ = p < 0.0005**

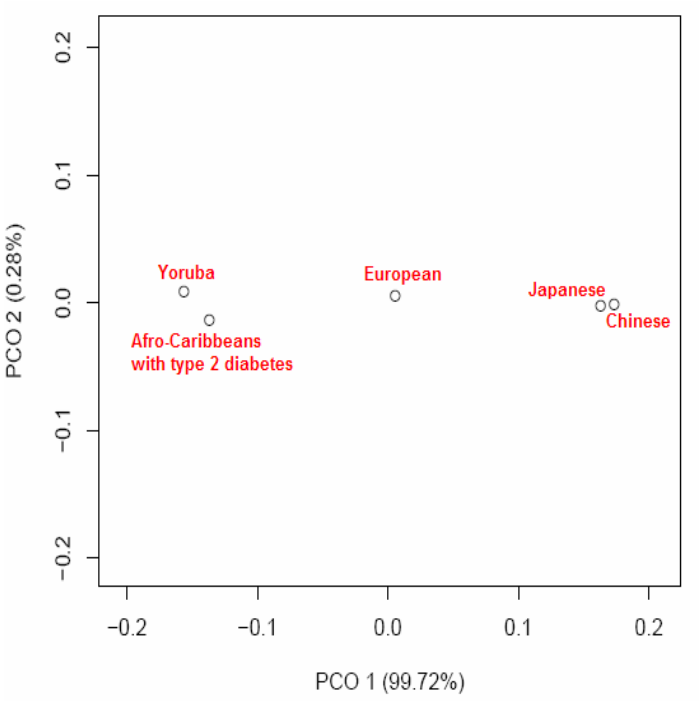
	Yoruba	European	Chinese	Japanese
Afro-Caribbbeans with T2D	-	+++	+++	+++

5.3.1.3 *TCF7L2* genetic structure in Afro-Caribbeans with T2D and HapMap populations

Hierarchical Fst (based on *TCF7L2* haplotypes) for Afro-Caribbeans and HapMap populations was 0.198 ($p < 0.00001$), with 19.8 % of the variation occurring among populations and 80.2 % occurring within populations.

A PCO plot of genetic distances (population pairwise Fsts are shown in supplementary figure S2) based on *TCF7L2* haplotypes is shown in figure 5.9. PCO 1 and 2 captured 100 % of the variation. Afro-Caribbeans with T2D lay between Yoruba and Europeans but were closest to Yoruba. East Asians formed a tight cluster furthest away from Afro-Caribbeans with T2D, and Europeans were almost equidistant between East Asians and Yoruba.

Figure 5.9 PCO plot of genetic distance (Fst) among Afro-Caribbeans with T2D and HapMap populations for *TCF7L2* haplotypes. Population pairwise Fsts are shown in supplementary figure S2.



5.3.1.4 Summary of results from Afro-Caribbeans with T2D and HapMap populations

In summary, these results show that whilst the distribution of *TCF7L2* haplotypes differed between Afro-Caribbeans with T2D and Europeans, Chinese and Japanese, against expectations, Yoruba were surprisingly similar to Afro-Caribbeans, even in terms of their TG at risk alleles. In accordance with their geographic proximity and recent East Asian ancestry, Chinese and Japanese groups were similar throughout the analysis. Consistent with their assumed mixed ancestry, Afro-Caribbeans were more similar to Yoruba and Europeans than East Asians. The *TCF7L2* haplotypes considered in this study were consequently capable of differentiating populations according to their recent ancestry and should prove useful data for

the study of inter-relationships and demographic history of Ethiopian populations. For this reason, and to place the results of this analysis into the context of Ethiopia, this study was extended to the Ethiopian ascertainment populations.

5.3.2 Variation in *TCF7L2* in the Ethiopian ascertainment populations

5.3.2.1 *TCF7L2* allele frequencies in the Ethiopian ascertainment populations

For each SNP, no population deviated significantly from Hardy Weinberg equilibrium at 5 % significance, except rs12255372 in Maale ($p = 0.004$). Following a Bonferroni correction for five tests, genotype distribution for Maale was not consistent with Hardy Weinberg equilibrium at the reduced p value threshold of 0.01. On closer inspection of rs12255372 genotypes, this deviation from Hardy Weinberg equilibrium was caused by an over representation of GG homozygotes in this group. All seven SNPs were common (≥ 8 %) in all groups, with many reaching frequencies > 20 % (figure 5.11).

5.3.2.2 LD in the Ethiopian ascertainment populations

Whilst the following intron 5 SNPs were in complete LD (D´ = 1) with each other in all groups: rs7901695, rs4132115, rs4506565, rs7068741, rs7069007, rs7069007 and rs7903146, many of the SNPs were not in complete LD with rs12255372 (intron 6). Varying levels of LD between rs7903146 and rs12255372 were observed in the Ethiopian ascertainment populations (figure 5.10). Similar to Europeans, LD was high in Amhara and Oromo. Similar to Yoruba, LD was low in Anuak.

Figure 5.10 LD between rs7903146 and rs12255372 in the Ethiopian ascertainment populations, Afro-Caribbeans with T2D and HapMap Yoruba, European and Japanese populations (data was uninformative with respect to LD in HapMap Chinese, as all samples were homozygous G for rs12255372)

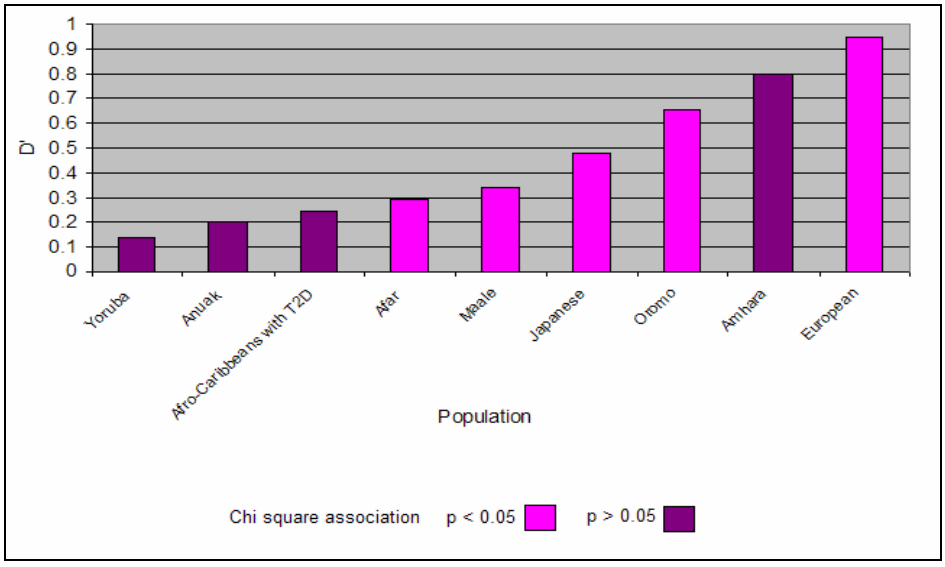


Figure 5.11 ***TCF7L2*** **SNP allele frequencies in the Ethiopian ascertainment populations** (derived alleles are always shown in purple)

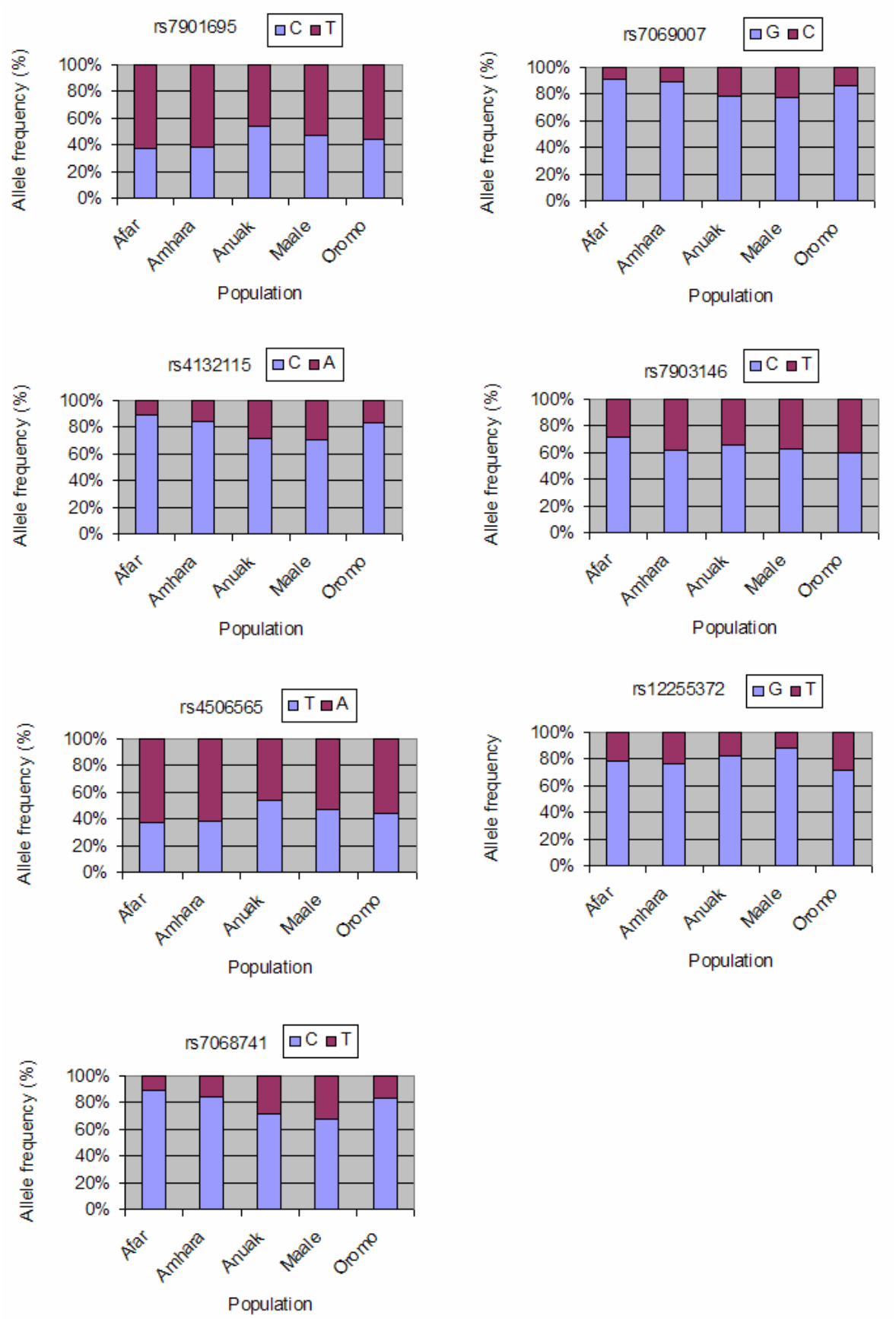
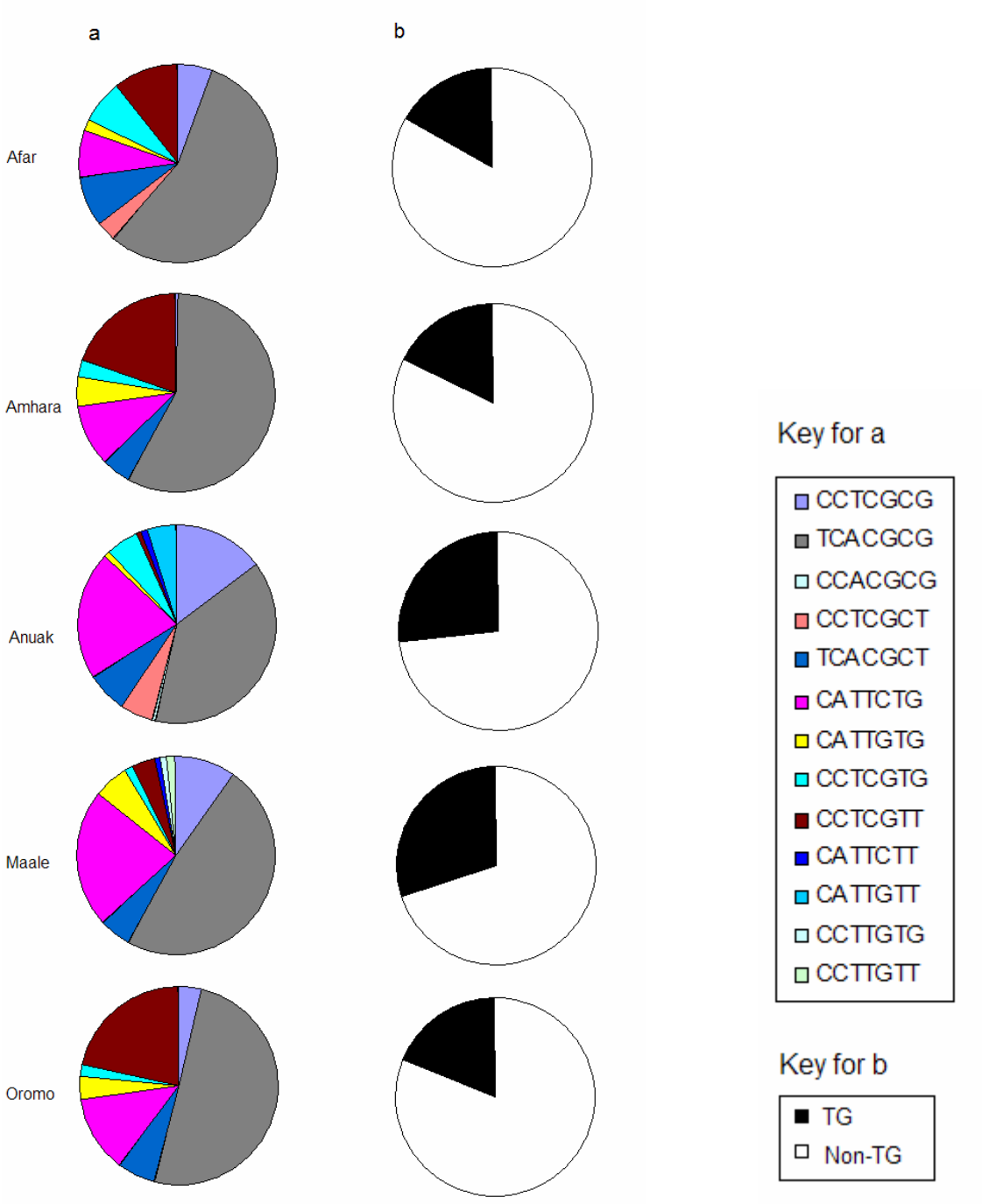


Figure 5.12 ***TCF7L2*** **haplotypes among the Ethiopian ascertainment populations.** Frequencies of haplotypes (rs7901695, rs4132115, rs4506565, rs7068741, rs7069007, rs7903146 and rs12255372) are shown in figure 5.12a. Proportions of haplotypes ending with rs7903146/rs12255372 TG (those associated with an increased risk of T2D in Afro-Caribbeans (Humphries et al., 2006)) are shown in figure 5.12b.



c Pairwise Fisher's exact tests comparing numbers of haplotypes 'ending with TG' versus those 'not ending with TG' shown above in (b) **- = p > 0.05**

	Afar	Amhara	Anuak	Maale	Oromo
Afro-Caribbeans with T2D	-	-	-	-	-

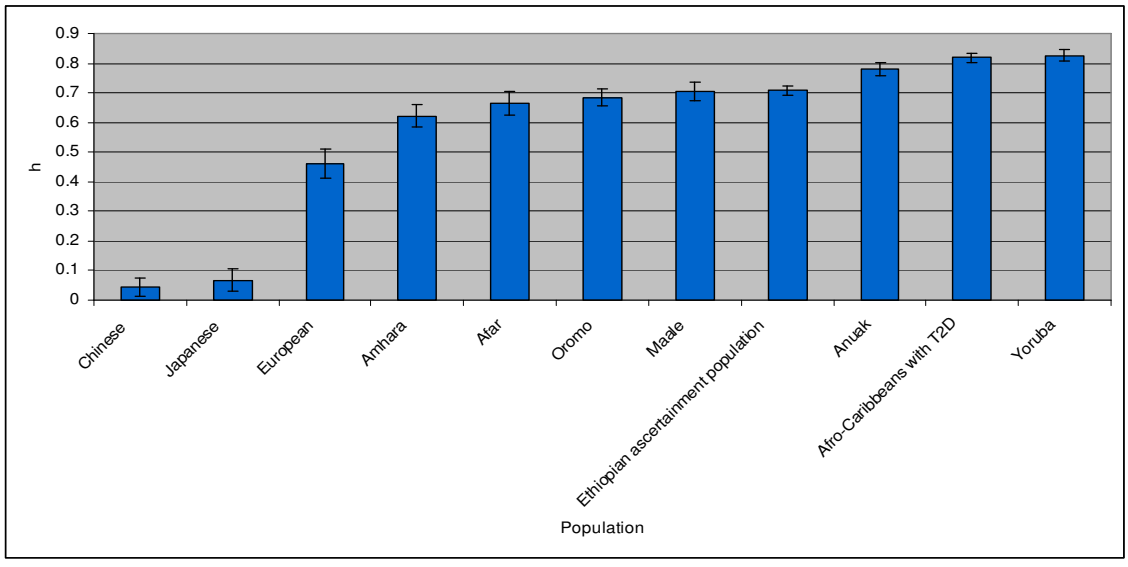
5.3.2.3 *TCF7L2* haplotypes in the Ethiopian ascertainment populations

A total of 13 haplotypes were observed in the Ethiopian ascertainment populations (figure 5.12a). Consistent with Afro-Caribbeans with T2D and HapMap populations (figure 5.8a), the most frequent haplotype for all populations was TCACGCG (≥ 38 % in any one group). Notably, all of the common variation observed in Afro-Caribbeans and HapMap populations (≥ 2 % in any one group) (figure 5.8a) was observed in the Ethiopian ascertainment populations (figure 5.12a). Consistent with Yoruba, none of the Ethiopian ascertainment populations were significantly different from Afro-Caribbeans with T2D in terms of their TG at risk alleles (figures 5.12b and 5.12c).

5.3.2.4 How diverse are the Ethiopian ascertainment populations?

Consistent with there being more human genetic diversity in Africans than in non-Africans, all African populations were more diverse than East Asians and Europeans (figure 5.13). Amongst the Africans, gene diversity was highest in Yoruba and Afro-Caribbeans with T2D and lowest in Amhara.

Figure 5.13 Gene diversity (h) based on *TCF7L2* haplotypes in various datasets (error bars represent standard deviation)



5.3.2.5 Can variation in *TCF7L2* differentiate populations?

For each SNP, Chinese and Japanese, and in some cases Europeans, were significantly different ($p < 0.05$) from most populations in terms of their genotypes (figure 5.14). Consistent with their recent African ancestry, the majority of African groups were not differentiated at 5 % significance (figure 5.14).

Figure 5.14 Exact test of population differentiation p values (lower triangle) and significant/not significant (+/-) differences at the 5 % threshold (upper triangle) for *TCF7L2* genotypes. All Afro-Caribbeans were type 2 diabetics.

rs7901695

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	+	-	-	+	+	-	+	+
Amhara	1.00		+	-	-	+	+	-	+	+
Anuak	0.01	0.02		-	-	-	-	+	+	+
Maale	0.25	0.35	0.29		-	-	-	+	+	+
Oromo	0.48	0.59	0.19	0.90		-	-	+	+	+
Afro-Caribbeans	0.02	0.02	0.35	0.52	0.21		-	+	+	+
Yoruba	0.01	0.01	0.08	0.18	0.08	0.68		+	+	+
European	0.30	0.25	p<0.01	0.02	0.04	p<0.01	p<0.01		+	+
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00	

rs4132115

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	+	+	-	+	-	+	+	+
Amhara	0.26		+	+	-	-	-	+	+	+
Anuak	p<0.01	0.02		-	-	-	+	+	+	+
Maale	p<0.01	0.01	0.88		+	-	+	+	+	+
Oromo	0.18	1.00	0.08	0.03		-	-	+	+	+
Afro-Caribbeans	p<0.01	0.19	0.27	0.32	0.28		-	+	+	+
Yoruba	0.05	0.54	0.03	0.04	0.47	0.48		+	+	+
European	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-	-
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00	1.00	

rs4506565

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	+	-	-	+	+	-	+	+
Amhara	1.00		+	-	-	+	+	-	+	+
Anuak	0.01	0.02		-	-	-	-	+	+	+
Maale	0.24	0.34	0.31		-	-	-	+	+	+
Oromo	0.48	0.59	0.18	0.91		-	-	+	+	+
Afro-Caribbeans	0.01	0.01	0.12	0.27	0.11		-	+	+	+
Yoruba	0.01	0.03	0.07	0.23	0.11	0.97		+	+	+
European	0.28	0.20	p<0.01	p<0.01	0.02	p<0.01	p<0.01		+	+
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00	

rs7068741

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	+	+	-	+	-	+	+	+
Amhara	0.25		+	+	-	-	-	+	+	+
Anuak	p<0.01	0.02		-	-	-	-	+	+	+
Maale	p<0.01	p<0.01	0.73		+	-	-	+	+	+
Oromo	0.19	1.00	0.07	0.01		-	-	+	+	+
Afro-Caribbeans	0.02	0.47	0.20	0.07	0.48		-	+	+	+
Yoruba	0.09	0.62	0.19	0.06	0.89	0.81		+	+	+
European	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-	-
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00	1.00	

rs7069007

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	+	+	-	-	-	+	+	+
Amhara	0.70		+	+	-	-	-	+	+	+
Anuak	p<0.01	p<0.01		-	+	-	-	+	+	+
Maale	p<0.01	p<0.01	1.00		+	-	-	+	+	+
Oromo	0.32	0.69	0.03	0.02		-	-	+	+	+
Afro-Caribbeans	0.10	0.29	0.12	0.09	0.80		-	+	+	+
Yoruba	0.14	0.41	0.10	0.06	0.83	1.00		+	+	+
European	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-	-
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00	1.00	

Figure 5.14 continued

rs7603146

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	-	-	-	-	-	-	+	+
Amhara	0.24		-	-	-	-	-	-	+	+
Anuak	0.51	0.85		-	-	-	-	-	+	+
Maale	0.24	1.00	0.90		-	-	-	-	+	+
Oromo	0.09	0.88	0.60	0.80		-	+	+	+	+
Afro-Caribbeans	0.46	0.82	0.92	0.83	0.52		-	-	+	+
Yoruba	0.28	0.09	0.12	0.08	0.04	0.25		-	+	+
European	0.81	0.09	0.25	0.12	0.04	0.18	0.09		+	+
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	1.00	

rs12255372

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		-	-	+	-	-	-	-	+	+
Amhara	0.32		-	+	-	-	-	-	+	+
Anuak	0.71	0.19		+	-	-	-	-	+	+
Maale	p<0.01	p<0.01	p<0.01		+	+	+	+	+	-
Oromo	0.37	0.48	0.10	p<0.01		-	-	-	+	+
Afro-Caribbeans	0.27	0.79	0.08	p<0.01	0.82		-	-	+	+
Yoruba	0.44	0.07	0.12	p<0.01	0.33	0.10		+	+	+
European	0.38	0.88	0.31	0.02	0.31	0.56	0.04		+	+
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-
Japanese	p<0.01	p<0.01	p<0.01	0.13	p<0.01	p<0.01	p<0.01	p<0.01	0.24	

When haplotypes using all seven SNPs (named seven SNP haplotypes hereafter) were analysed (figure 5.15a), the majority of populations were significantly different at 5 % significance. Consistent with language similarity, Afar and Oromo, both speaking languages in the Cushitic family, were not significantly different.

Figure 5.15 Exact test of population differentiation p values (lower triangle) and significant/not significant (+/-) differences at the 5 % threshold (upper triangle) for *TCF7L2* haplotypes. All Afro-Caribbeans were type 2 diabetics.

a) Based on haplotypes from all seven SNPs

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		+	+	+	+	+	+	+	+	+
Amhara	p<0.01		+	+	-	+	+	+	+	+
Anuak	p<0.01	p<0.01		+	+	+	+	+	+	+
Maale	p<0.01	p<0.01	p<0.01		+	+	+	+	+	+
Oromo	p<0.01	0.47	p<0.01	p<0.01		+	+	+	+	+
Afro-Caribbeans	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-	+	+	+
Yoruba	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	0.46		+	+	+
European	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		+	+
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	0.63	

b) Based on haplotypes from rs7903146 (intron 5) and rs12255372 (intron 6)

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar		+	-	+	+	+	+	+	+	+
Amhara	0.02		+	+	-	+	+	+	+	+
Anuak	0.12	p<0.01		-	+	-	-	+	+	+
Maale	p<0.01	p<0.01	0.26		+	+	+	+	+	+
Oromo	0.03	0.90	p<0.01	p<0.01		+	+	+	+	+
Afro-Caribbeans	0.03	p<0.01	0.22	p<0.01	p<0.01		-	+	+	+
Yoruba	p<0.01	p<0.01	0.17	p<0.01	p<0.01	0.68		+	+	+
European	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		+	+
Chinese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01		-
Japanese	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	0.62	

When haplotypes using only two SNPs (rs7903146 in intron 5 and rs12255372 in intron 6 and named two SNP haplotypes hereafter) were analysed (figure 5.15b), the majority of information observed from seven SNP haplotypes was captured. Only four population pairs were no longer differentiated at 5 % significance. Consistent with Anuak having Y chromosome types similar to those prevalent in West Africa but not widely observed elsewhere in Ethiopia (unpublished data), Anuak were not significantly different from Yoruba or Afro-Caribbeans. In accordance with their geographical proximity, Anuak and Maale, both residing to the west of Ethiopia, were not significantly different. Despite living in very different areas of the country, Afar (NE) and Anuak (W) were also not significantly different.

When the pattern of *TCF7L2* genetic structuring within and among various populations in different datasets was analysed (table 5.2), similar results were observed from seven SNP haplotypes and two SNP haplotypes. In all datasets, the vast majority (≥ 90 %) of variation was observed within populations.

Table 5.2 Hierarchical Fst based on seven SNP haplotypes (7) and two SNP haplotypes (2) in various datasets. All Fsts were significant ($p < 0.00001$).

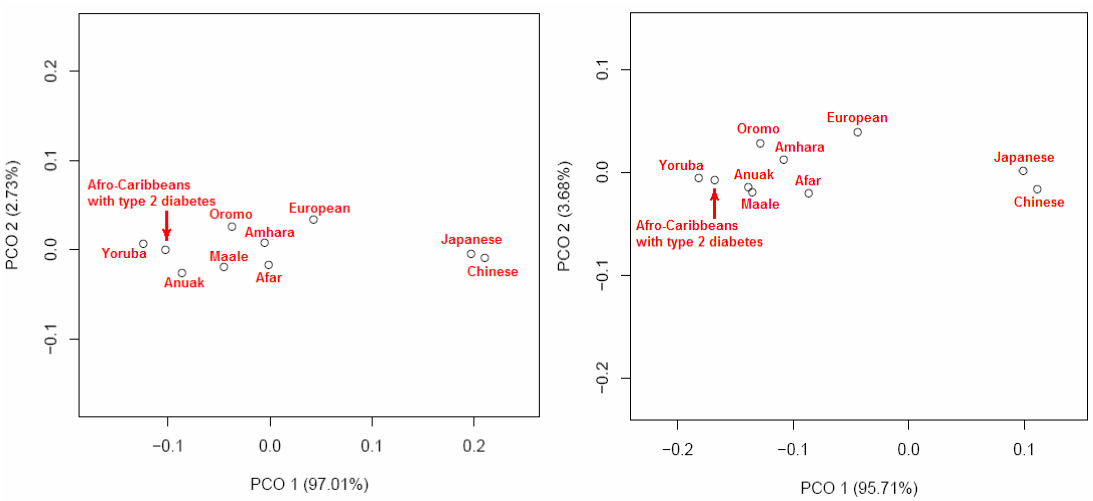
Dataset	Fst		Variation among populations (%)		Variation within populations (%)	
	7	2	7	2	7	2
Individual Ethiopian ascertainment populations, Afro-Caribbeans and individual HapMap populations	0.09	0.08	9.28	7.81	90.72	92.19
Pooled Ethiopian ascertainment population, Afro-Caribbeans and individual HapMap populations	0.11	0.29	10.59	9.65	89.41	90.35
Individual Ethiopian ascertainment populations	0.03	0.01	2.74	1.29	97.26	98.71

A similar pattern was observed when population pairwise genetic distances (Fsts) based on seven SNP and two SNP haplotypes were analysed (figure 5.16). Afro-Caribbeans lay between Yoruba and Europeans but were closest to Yoruba. Notably, of all the groups, Yoruba were always closest to Afro-Caribbeans with T2D. East Asians formed a tight cluster furthest away from the other groups. Anuak lay closest to Yoruba and Afro-Caribbeans. Consistent with their assumed mixed ancestry, Afar, Amhara and Oromo clustered almost equidistantly between Yoruba and Europeans, and Amhara were the closest Ethiopian group to Europeans.

Given their capacity to differentiate the Ethiopian ascertainment populations, rs7903146 and rs12255372 genotypes were obtained for over 6000 individuals from multiple Ethiopian ethnic groups. In addition to providing insight into T2D susceptibility in Ethiopians, this data should prove useful for the study of inter-relationships and demographic history of Ethiopian populations.

Figure 5.16 PCO plots of genetic distance (Fst) among various populations for *TCF7L2* haplotypes. Population pairwise Fsts are shown in supplementary figure S3. Both plots are largely similar.

a) Based on seven SNP haplotypes **b) Based on two SNP haplotypes**



5.3.3 Variation in *TCF7L2* in multiple Ethiopian populations

5.3.3.1 rs7903146 and rs12255372 allele frequencies in multiple Ethiopian populations

For each SNP, no population deviated significantly from Hardy Weinberg equilibrium at 5 % significance, except rs7903146 in Mashile (p = 0.01) and rs12255372 in Dasenech (p = 0.02), Maale (p = 0.04) and Yem (p = 0.01). In view of multiple testing however, a p value of 0.05 indicated no significant deviation from Hardy Weinberg; with a Bonferroni correction for 50 tests, a p value of less than 0.001 was considered significant. Genotype distributions for the above groups were therefore consistent with Hardy Weinberg equilibrium at this reduced p value threshold.

rs7903146 T allele frequencies were very common in all populations (figure 5.17) ranging from 0.24 (n = 13) in Dasenech to 0.56 (n = 259) in Ari cultivators, and reaching ≥ 0.40 in most groups. For the pooled Ethiopian dataset, rs7903146 T was observed at a frequency of 0.41 (n = 5246).

rs12255372 T allele frequencies were also observed in all populations (figure 5.18) but were less common than rs7903146 T. Allele frequencies ranged from 0.01 (n = 1) in Ari smiths to 0.26 (n = 76) in Oromo, and reached ≥ 0.10 in most groups. For the pooled Ethiopian dataset, rs12255372 T was observed at a frequency of 0.14 (n = 1821).

Figure 5.17 rs7903146 allele frequencies in multiple Ethiopian populations

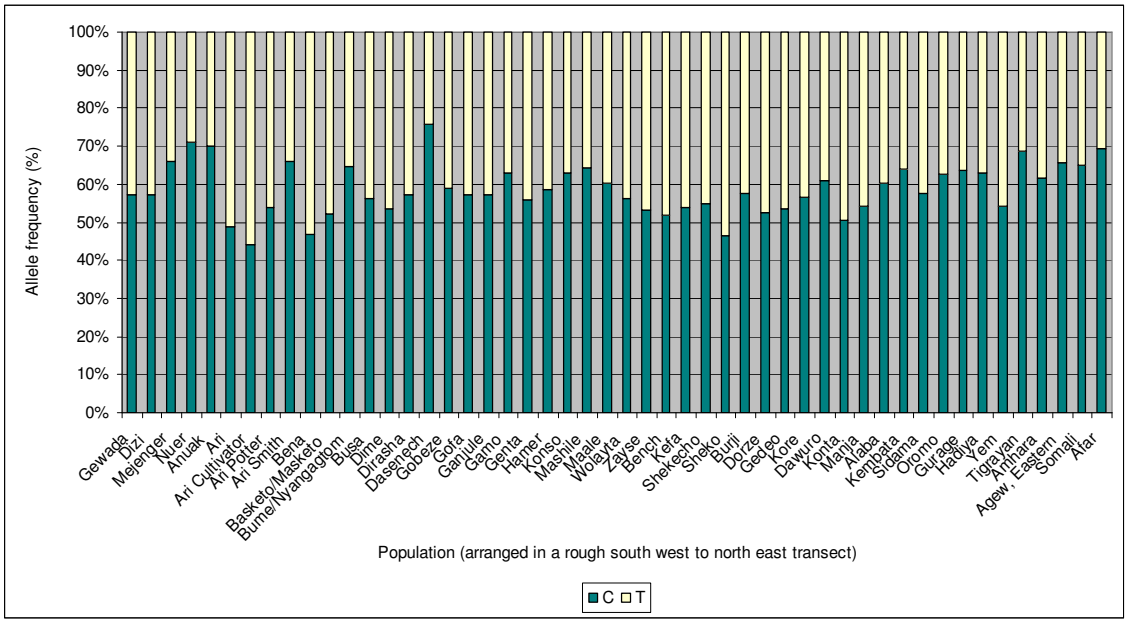
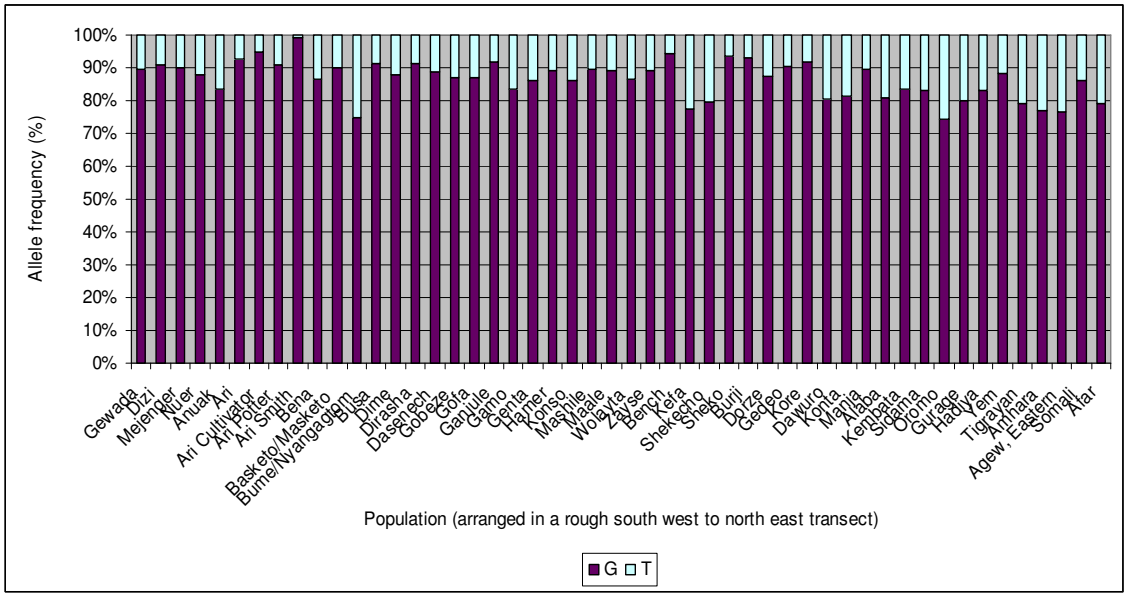


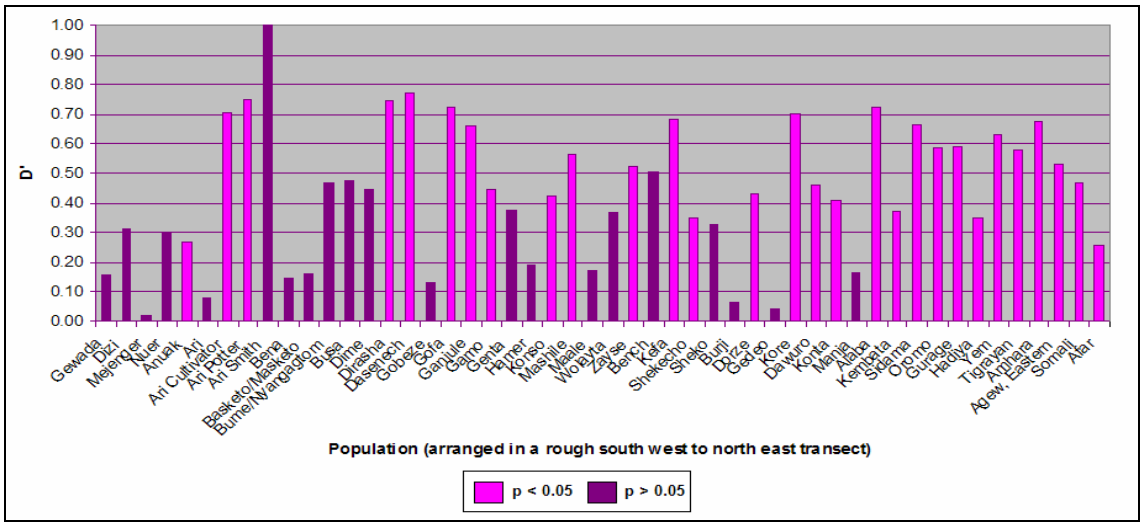
Figure 5.18 rs12255372 allele frequencies in multiple Ethiopian populations



5.3.3.2 LD between rs7903146 and rs12255372 in multiple Ethiopian populations

LD between both SNPs was moderately high for the pooled Ethiopian dataset (D´ = 0.44, Chi square association $p < 0.01$). LD varied among the different Ethiopian populations (figure 5.19). Complete LD (D´ = 1) was only observed in Ari smiths. Where D´ was not equal to 1, LD ranged from 0.02 in Mejenger to 0.77 in Dasenech, and was ≥ 0.40 in most groups. LD was generally higher in groups towards the north east of Ethiopia and lower in groups to the south west.

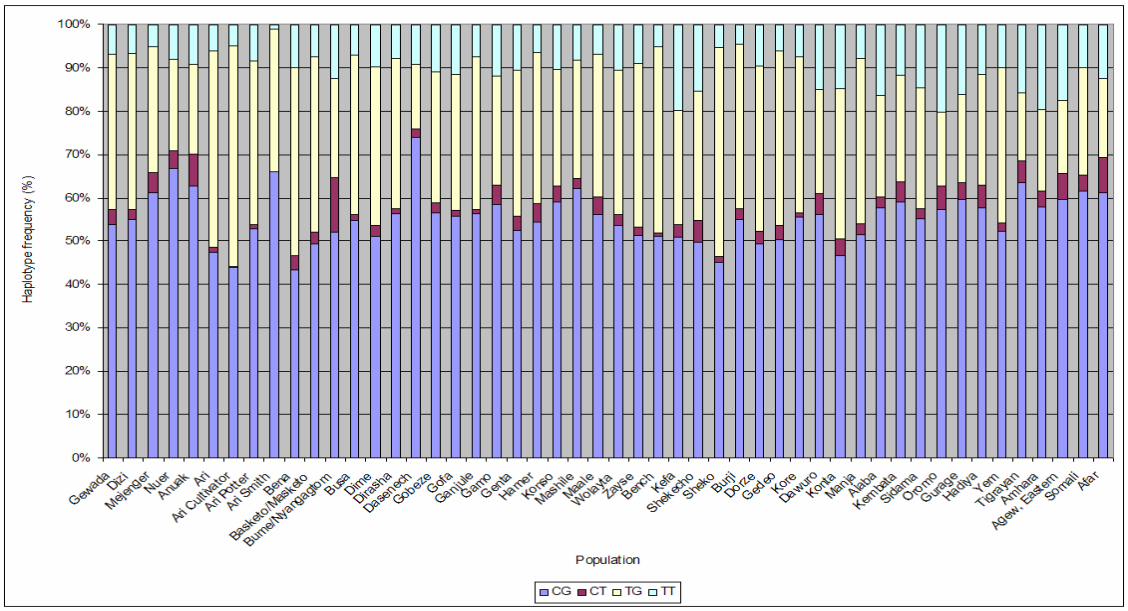
Figure 5.19 LD between rs7903146 and rs12255372 in multiple Ethiopian populations. P values are for Chi square associations.



5.3.3.3 rs7903146/rs12255372 haplotype frequencies in multiple Ethiopian populations

For the pooled Ethiopian dataset, CG was the modal haplotype (55 %, n = 7009). The T2D at risk haplotypes (TG and TT (Humphries et al., 2006)) were common (30 %, n = 3831 and 11 %, n = 1400 respectively) and CT was observed at a frequency of 3 % (n = 412). Consistent with D´ values, all four haplotypes were reported in all populations but one (CT was not observed in Ari smiths) (figure 5.20). CG was the modal haplotype (≥ 47 %) in most groups. TG was generally the next most frequent haplotype and reached ≥ 30 % in most groups. TT was generally the third most frequent haplotype, reaching ≥ 10 % in most groups. CT was generally the least frequent haplotype, reaching ≥ 3 % in most groups.

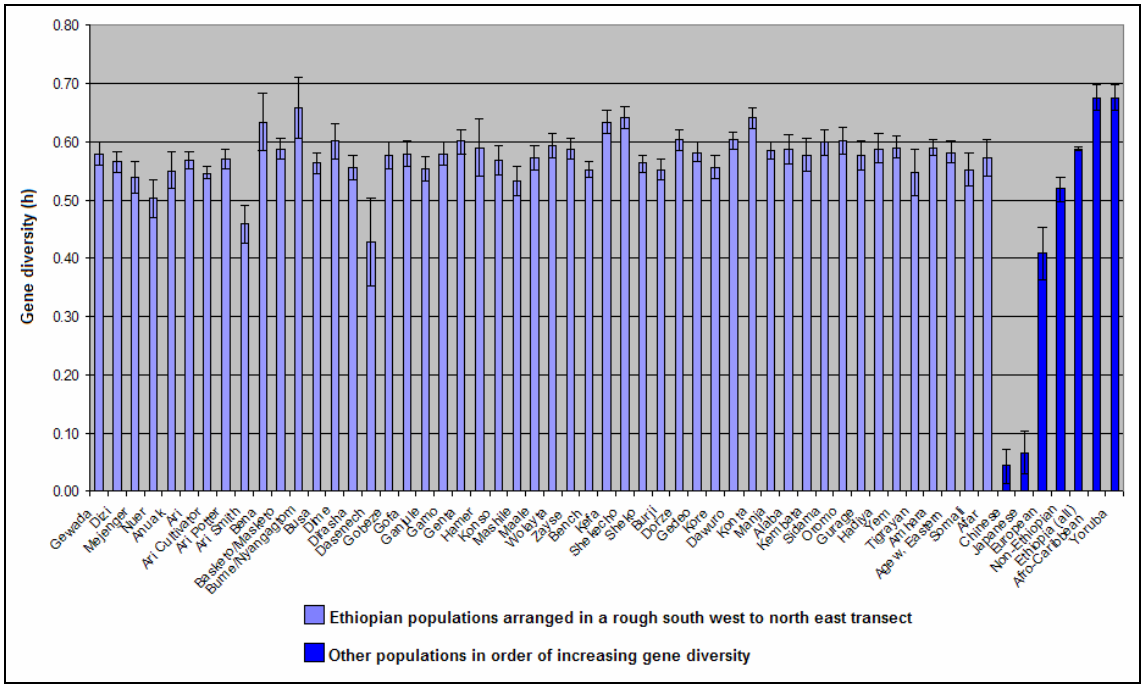
Figure 5.20 rs7903146/rs12255372 haplotype frequencies in multiple Ethiopian populations



5.3.3.4 Gene diversity for rs7903146/rs12255372 haplotypes in multiple Ethiopian populations

Gene diversity for the Ethiopian populations (light blue bars in figure 5.21) was moderately high in all groups (≥ 0.43) and ranged from 0.43 ± 0.07 in Dasenech to 0.66 ± 0.05 in Bume/Nyangagtom, with a mean of 0.57 and a variance of 0.002. In the context of gene diversity in other populations (dark blue bars in figure 5.21), all Ethiopian groups lay between Europeans and both Yoruba and Afro-Caribbeans with T2D, and Dasenech were most similar to Europeans.

Figure 5.21 Gene diversity for rs7903146/rs12255372 haplotypes in multiple Ethiopian populations. Error bars represent standard deviation.



5.3.3.5 How different are the Ethiopian populations in terms of their rs7903146/rs12255372 haplotype frequencies?

5.3.3.5.1 Population differentiation

In accordance with their recent non-African ancestries, HapMap Europeans, Chinese and Japanese populations were significantly different ($p > 0.05$) from all African groups (figure 5.22). Consistent with geography, Hamer, Bena and Bume/Nyangagtom, who live in the south west of Ethiopia, were not different from Yoruba at 5 % significance. Notably, Hamer, Bena and Bume/Nyangagtom were the only Ethiopian populations to not significantly differ ($p < 0.05$) from Afro-Caribbeans with T2D. Consistent with there being substantial human genetic variation in Ethiopia, 53 % ($n = 646$) of pairwise comparisons among Ethiopian populations were significantly different ($p < 0.05$). In accordance with their marginalised societies, Manja and

Figure 5.22 Exact test of population differentiation significant/not significant (+/-) differences at the 5% threshold for rs7903146/rs12255372 haplotypes. Afro-Caribbeans have T2D. Bume are also known as Nyangagtom, and Basketo are also known as Masketo.

	Ari	Ari Cultivator	Afar	Agew, Eastern	Alaba	Amhara	Anuak	Ari Potter	Ari Smith	Bena	Basketo	Bume	Bench	Burji	Busa	Dizi	Dime	Dirasha	Dasenech	Dawuro	Dorze	Gobeze	Gedeo	Gofa	Ganjule	Gamo	Genta	Gurage	Gewada	Hadiya	Hamer	Kefa	Kembata	Konta	Kore	Konso	Manja	Mejenger	Mashile	Maale	Nuer	Oromo	Shekecho	Sidama	Sheko	Somali	Tigrayan	Wolayta	Yem	Zayse	Afro-Caribbean	Yoruba	European	Chinese
Ari Cultivator	-																																																				
Afar	+	+																																																				
Agew, Eastern	+	+	-																																																		
Alaba	+	+	+	+																																																
Amhara	+	+	+	-	-																																															
Anuak	+	+	-	+	+	+																																													
Ari Potter	-	+	+	+	+	+	+																																												
Ari Smith	+	+	+	+	+	+	+	+																																											
Bena	-	-	+	+	-	+	-	-	+																																									
Basketo	-	+	+	+	+	+	+	-	+	-																																								
Bume	+	+	-	-	+	+	-	+	+	-	+																																								
Bench	-	-	+	+	+	+	+	-	+	-	-	+																																							
Burji	-	+	+	+	+	+	+	-	-	-	-	+	-																																						
Busa	-	+	+	+	+	+	+	-	+	-	-	+	-	-																																					
Dizi	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-																																			
Dime	-	+	+	+	-	+	+	-	+	-	-	-	-	-	-	-																																		
Dirasha	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-																																
Dasenech	+	+	-	-	-	-	-	+	+	+	+	-	+	+	+	+	+	+																															
Dawuro	+	+	+	+	-	+	+	+	+	-	+	-	+	+	+	+	-	+	-																														
Dorze	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+																													
Gobeze	+	+	+	+	-	+	+	-	+	-	-	+	+	+	-	-	-	-	-	+	-																												
Gedeo	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	+																										
Gofa	+	+	+	+	-	+	+	-	+	-	-	+	+	+	-	-	-	-	+	+	-	-	+																									
Ganjule	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-																							
Gamo	+	+	-	+	-	+	-	+	+	-	+	-	+	+	+	+	-	+	-	-	+	-	+	-	+																						
Genta	+	+	+	+	-	+	+	-	+	-	-	+	+	-	-	-	-	-	+	+	-	-	-	-	-	-																					
Gurage	+	+	-	-	-	-	+	+	+	+	+	-	+	+	+	+	+	+	-	-	+	+	+	+	+	-	+																			
Gewada	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+																		
Hadiya	+	+	-	+	-	+	-	+	+	-	+	-	+	+	+	+	-	+	-	-	+	-	+	-	+	-	-	-	+																	
Hamer	-	+	-	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
Kefa	+	+	+	+	-	-	+	+	+	-	+	+	+	+	+	+	-	+	+	-	+	-	+	-	+	+	+	-	+	+	-															
Kembata	+	+	-	+	-	+	-	+	+	-	+	-	+	+	+	+	-	+	-	-	+	-	+	-	+	-	-	-	+	-	-	+														
Konta	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+	-	+	+	+	-	-	+	-	+	+	-	+	-	+	-	-	+													
Kore	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+	-	+	-	+	+	+												
Konso	+	+	+	+	-	+	-	+	+	-	+	-	+	+	+	-	-	-	-	-	-	-	+	-	-	-	-	+	-	+	-	+	+	-	+											
Manja	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+	-	+	-	+	+	-	-	+										
Mejenger	+	+	+	+	+	+	-	+	+	-	+	+	+	-	+	-	-	+	-	+	+	-	-	+	+	+	+	+	-	-	-	+	-	+	+	-	+									
Mashile	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	-	+	+	-	+	-	-	-	-	+	-	-	-	+	-	+	-	-	+	-								
Maale	+	+	+	+	+	+	+	-	+	-	-	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	+	-	-	-	+	-	+	-	-	-	-	-							
Nuer	+	+	-	+	+	+	-	+	+	-	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	+	-	-	-	+	-	+	+	-	-	-	-	+						
Oromo	+	+	-	-	-	-	+	+	+	+	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+					
Shekecho	+	+	+	+	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	-	-	+	-	+	-	-	+	+	-	-	-	-	-	+	-	+	+	+	+	+	+				
Sidama	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	+	-	-	-	-	-	-	+	-	+	-	-	-	+	-	-	-	-	-	+	-	+	+	-	+	+	+	-			
Sheko	-	-	+	+	+	+	+	-	+	-	-	+	-	-	-	+	-	+	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
Somali	+	+	-	+	-	+	-	+	+	-	+	-	+	+	-	-	-	-	-	-	+	-	+	-	+	-	-	-	+	-	-	-	-	+	-	+	+	-	+	+	-	-	+			
Tigrayan	+	+	-	-	-	-	-	+	+	+	+	-	+	+	+	+	+	+	-	-	+	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+		
Wolayta	+	+	+	+	-	+	+	-	+	-	-	+	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	+	+	-	-	+	-	+				
Yem	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+	-	+	-	+	+	-	-	-	-	+	-	-	+	-	+			
Zayse	-	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+	-	+	-	+	+	-	-	-	-	+	-	-	+	-	+	-		
Afro-Caribbean	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
Yoruba	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
European	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Chinese	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Japanese	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-

Dawuro were significantly different ($p < 0.05$) and Ari cultivators differed from Ari potters and smiths at 5 % significance. Ari potters and smiths were also significantly different ($p < 0.05$). Notably, Ari (dataset not considering Ari social structure and collected prior to the collection of the Ari cultivators, potters and smiths) were not significantly different from Ari cultivators or potters but were different from the Ari smiths at 5 % significance.

5.3.3.5.2 Genetic structure

Hierarchical Fst values (table 5.3) showed that, whilst most variation was observed within populations, inter-population variation was evident, with 2 % of variation being observed among Ethiopian populations.

Table 5.3 Hierarchical Fst based on rs7903146/rs12255372 haplotypes in various datasets. All Fsts were significant ($p < 0.00001$).

Dataset	Fst	Variation among populations (%)	Variation within populations (%)
Individual Ethiopian populations	0.02	1.98	98.03
Ethiopian populations plus HapMap Yoruba, Europeans, Chinese and Japanese	0.03	2.78	97.22
Pooled Ethiopian population plus HapMap Yoruba, Europeans, Chinese and Japanese	0.32	8.89	91.11

5.3.3.5.3 Genetic distance

A PCO plot of population pairwise Fsts (figure 5.23a) showed that the vast majority of Ethiopians cluster together. Afro-Caribbeans with T2D tightly cluster with Yoruba and lie close to the Ethiopian cluster. Chinese and Japanese are close but furthest away from the other populations. Europeans are not tightly clustered with any population but are closest to Dasenech. At closer resolution (figure 5.23b), Yoruba remain close to Afro-Caribbeans with T2D.

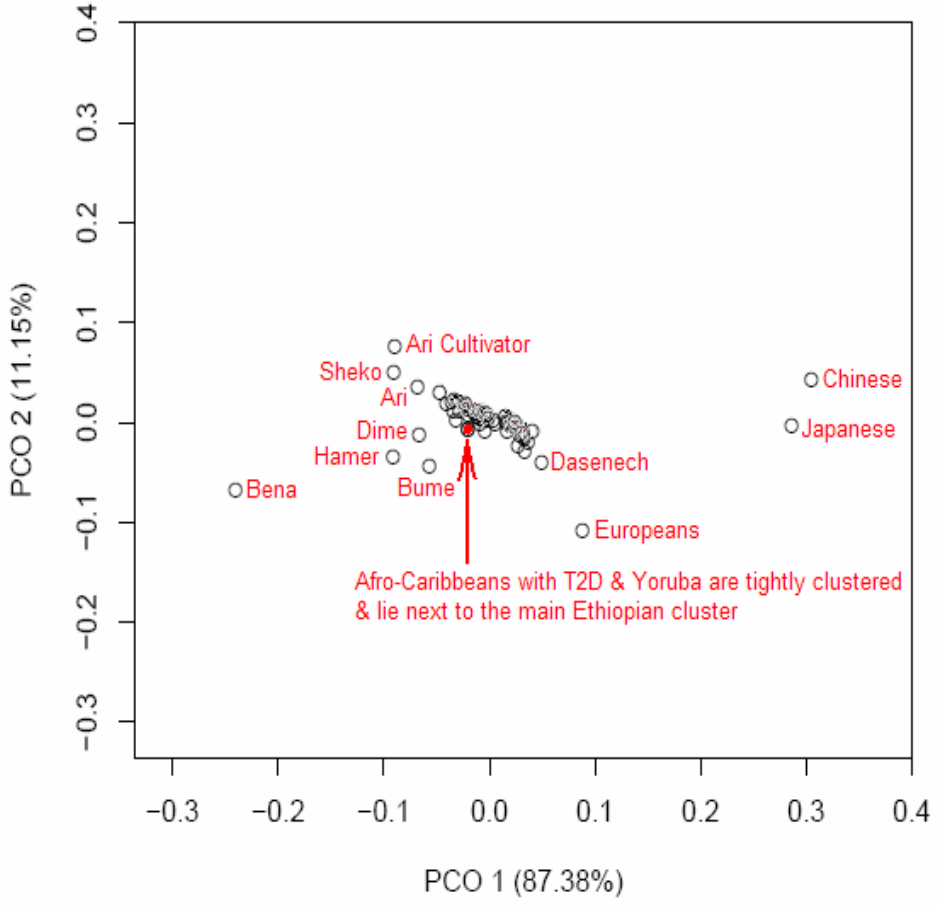
5.3.3.5.4 Is genetic distance correlated with geography, language and/or religion?

Pairwise genetic distances (Fst), based on rs7903146/rs12255372 haplotypes in the Ethiopian populations (supplementary figure S4), were correlated with pairwise average geographic distances (supplementary figure S5) (mantel z statistic = 6443.91, $p < 0.01$). Given the correlation between genetic distance and geographic distance:

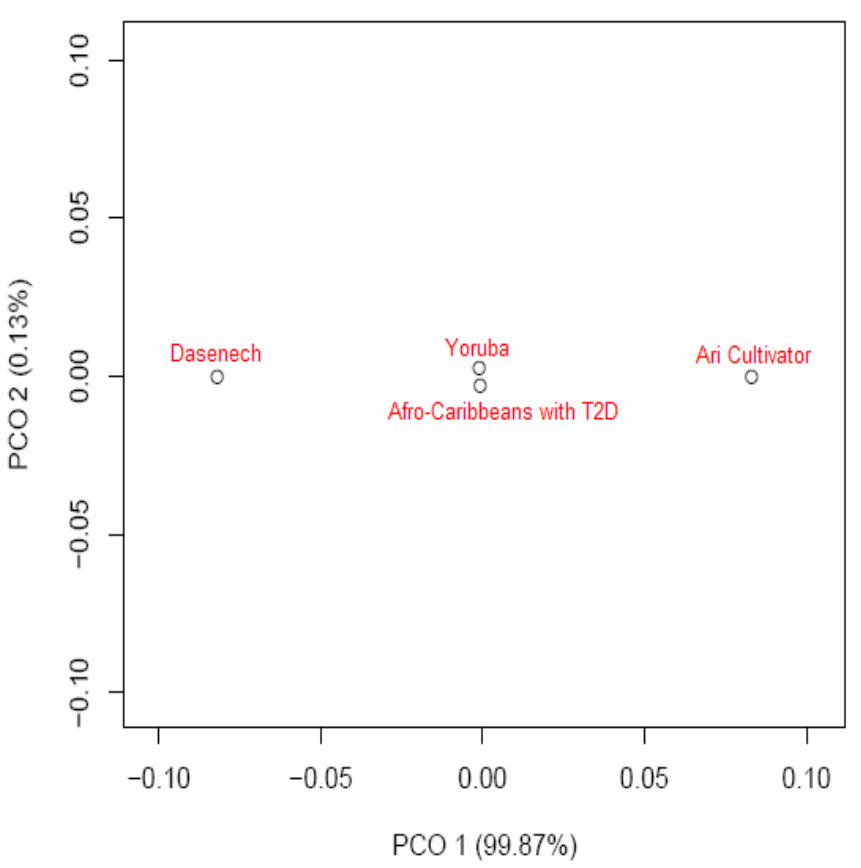
- Pairwise language distances (Fst), based on donor's first language (supplementary figure S6), were not correlated with pairwise genetic distances (partial mantel statistic $r = 0.06$, $p > 0.1$).

Figure 5.23 PCO plots of pairwise genetic distances (Fst) based on rs7903146/rs12255372 haplotypes. Pairwise Fsts are shown in supplementary figure S4 (the majority (57 %) of pairwise Fst values were significant at the 5 % significance threshold). For each plot, PCOs 1 and 2 capture > 98 % of the variation.

a) All populations: The majority of Ethiopian populations cluster together away from Europeans & East Asians. Yoruba are closest to Afro-Caribbeans with T2D.



b) At closer resolution: Yoruba remain closest to Afro-Caribbeans with T2D.



- Pairwise language distances (Fst), based on linguistic group in which donor's first language falls (supplementary figure S7), were weakly correlated with pairwise genetic distances (partial mantel statistic $r = 0.20$, $p < 0.001$).

- Pairwise religious distances (Fst) (supplementary figure S8) were not correlated with pairwise genetic distances (partial mantel statistic $r = -0.11$, $p > 0.1$).

5.4 Discussion

5.4.1 Variation in *TCF7L2* in Afro-Caribbeans with T2D

Since the causative variant of the association of *TCF7L2* with an increased risk of T2D is not known, it was important to further analyse TG rs7903146/rs12255372 haplotypes which are thought to be associated with T2D susceptibility in a type 2 diabetic Afro-Caribbean cohort from Humphries et al. (2006). By extending the TG rs7903146/rs12255372 haplotype to include SNPs found in LD block 5 of HapMap Yoruba (rs7901695, rs4132115, rs4506565, rs7068741 and rs7069007), this study aimed to investigate whether a specific TG haplotype might be a potential T2D risk haplotype in Afro-Caribbeans with T2D. Ethnically matched non-diabetic controls were not available for this study, and as a consequence, variability in haplotype distribution due to differential demographic histories was not controlled. HapMap populations were used as controls, and it was assumed that each control population was representative of their respective general populations, with some individuals being prone to T2D and some not. Since Afro-Caribbeans are thought to be an admixed population with recent European and West African ancestry (Miljovic-Gacic et al., 2005), HapMap European and Yoruba (West African ethnic group) populations were useful as controls in this respect.

Results showed that whilst haplotype distribution differed between Afro-Caribbeans with T2D and Europeans, Chinese and Japanese groups, Yoruba were always strikingly similar to Afro-Caribbeans, even in terms of their 'TG at risk' alleles. Of the rs7903146/rs12255372 haplotypes in Afro-Caribbeans with T2D, Humphries et al. (2006) only reported an association between TG and an increased risk of the disease. The association was not significant ($p = 0.24$) and the study suggested that this may be due to small sample size, but the similar distribution of TG haplotypes in both Afro-Caribbeans with T2D and Yoruba suggests that the association may be due to a demographic effect. Since a high frequency of TG haplotypes were identified in Yoruba (and may be identified in other West Africans), further studies are warranted to clarify whether the TG haplotype is significantly associated with an increased risk of the disease or not. If the association is confirmed, the results from the study presented here suggest that *TCF7L2* genotypes may confer the same T2D risk in both Afro-Caribbeans and Yoruba.

It is important to note that while the Afro-Caribbeans in this study were not a random sample of the general Afro-Caribbean population, they were a random sample of Afro-Caribbeans with

T2D. The inclusion of this sample in the population level analyses performed throughout this study consequently obeys the 'random sampling' assumption of some population genetics statistical tests, e.g. Fst and exact test of population differentiation (Schneider et al., 2000). Including a random sample of the general Afro-Caribbean population would however place data from this study into the context of Afro-Caribbeans without T2D. The similarity of Yoruba to the type 2 diabetic Afro-Caribbean cohort throughout this study does not necessarily mean that Yoruba may be more susceptible to T2D than the general Afro-Caribbean population.

5.4.2 Placing the *TCF7L2* data into the context of Ethiopia

Genotypes for each of the seven *TCF7L2* SNPs analysed in this study failed to differentiate the majority of African groups (Ethiopian ascertainment populations, Yoruba and Afro-Caribbeans with T2D). Many differences were however observed among groups when haplotypes were considered. Furthermore, two SNP haplotypes (based on rs7903146 and rs12255372) captured the majority of the information on intra and inter-population diversity from haplotypes of all seven SNPs. Given that LD between rs7903146 and rs12255372 varied among groups and was generally low in Africans, it is evident that more information is available from haplotype data than genotype data when analysing populations with low overall levels of LD between two markers.

In this study, African populations were more diverse than non-Africans. This is consistent with anatomically modern human originating in Africa. Furthermore, all of the common variation observed outside Africa was observed in the Ethiopian ascertainment populations. This is consistent with anatomically modern humans expanding out of Africa via Ethiopia ~ 70 thousand years ago (Wallace et al., 1999), and a more recent migration of Semitic speaking peoples from Arabia to Ethiopia.

The 'at risk' TG and TT haplotypes were common throughout Ethiopia suggesting that substantial numbers of Ethiopians might be prone to T2D (it is however noteworthy to mention that T2D is a complex disease and many factors, both genetic and environmental, contribute to an individual's T2D susceptibility). These findings are not only of benefit to the indigenous populations of Ethiopia, but are also of increasing importance in directing public healthcare policies in the developed world, where growing numbers of individuals with recent Ethiopian descent are adopting Western dietary habits which may exacerbate the onset of T2D.

Inter-population variation was observed in Ethiopians, and the majority of Ethiopian populations were significantly different in terms of their rs7903146/rs12255372 haplotypes, suggesting that Ethiopians should not be treated as one group, but as several different groups. In most Ethiopian populations, the TG 'at risk' haplotype (known to be common in Afro-Caribbeans (Humphries et al., 2006)) was generally observed at higher frequencies than the TT 'at risk' haplotype (known to be common in Europeans (Humphries et al., 2006)). Furthermore,

consistent with their assumed mixed ancestry from the Arabic north and sub-Saharan Africa (Cavalli-Sforza et al., 1994), Ethiopian populations lay between Yoruba and Europeans, but closer to Yoruba, in PCO plots of genetic distance. These results suggest that most populations may have more of a West African *TCF7L2* risk for T2D, with few populations possibly having more of a European *TCF7L2* risk.

Consistent with the isolation by distance model which predicts that population similarity will decline with increasing geographical distance as a result of the limiting effect of geographic distance on rates of gene flow (Hardy and Vekemans, 1999), pairwise genetic distance (based on two SNP haplotypes) was correlated with geographic distance between multiple Ethiopian populations. Given this correlation, genetic distance was also correlated with linguistic differences. It is apparent that, for these markers in this gene, genes do not travel according to group specific language or religion. These results are largely consistent with Y chromsome data which shows an inter-ethnic division between the north Semitic and south non-Semitic speaking groups in Ethiopia (unpublished data). This study consequently highlights the power of using two SNP haplotypes when analysing populations with low overall levels of LD between the two markers.

LD between rs7903146 and rs12255372 was generally higher in groups towards the north east of Ethiopia and lower in groups to the south west. Populations with varying levels of LD may prove useful in T2D genotype/phenotype association studies in the future. They may also be informative in association studies investigating the role of rs7903146 in the aetiology of T2D.

In accordance with their hierarchical societies, Manja and Dawuro were significantly different as were Ari cultivators, Ari potters and Ari smiths. Ari (dataset not considering Ari social structure and collected prior to the collection of Ari cultivators, potters and smiths) were not different from Ari cultivators or Ari potters but were different from Ari smiths, suggesting that the Ari collection may have excluded Ari smiths. These results highlight the importance of considering group social structure when collecting samples from ethnic groups in Ethiopia, not least in southern Ethiopia where hierarchical societies are common place (Freeman and Pankhurst, 2003).

5.5 Conclusion

In the course of investigating whether a specific TG rs7903146/rs12255372 haplotype might be a potential T2D risk haplotype in Afro-Caribbeans with T2D, this study found that *TCF7L2* might confer the same risk of T2D in both Yoruba and Afro-Caribbeans. Depending on variation in other genes and environmental factors, this in turn suggests that Yoruba, and potentially other West Africans, might be as prone to T2D as Afro-Caribbeans. With particular relevance for healthcare in Ethiopia, appreciable numbers of Ethiopians might be susceptible to T2D, with most populations having more TG 'at risk' haplotypes, hence more of a West African *TCF7L2* risk for the disease, and some having more TT 'at risk' haplotypes, hence more of a European

TCF7L2 risk. Since LD between rs7903146 and rs12255372 varied among Ethiopians, studies involving Ethiopians might prove particularly informative in T2D genotype/phenotype association studies, and in association studies investigating the causal variant of the association of *TCF7L2* with predisposition to the disease. With particular relevance for anthropological studies, this study also found that the rs7903146/rs12255372 haplotypes could effectively discriminate Ethiopian populations in accordance with geography and linguistics, providing insight into inter-population relationships and population histories. The haplotypes even had enough power to differentiate caste-like groups covered by the same self-identifying ethnic label. Consistent with the hypothesis that anatomically modern humans expanded out of Africa and into the New World via Ethiopia, all of the common *TCF7L2* variation observed in non-African populations was observed in Ethiopia.

6 General discussion

According to the Out of Africa model, anatomically modern humans originated in Africa ~ 200 thousand years ago and colonised the rest of the world within the last ~ 100 thousand years (Stringer and Andrews, 1988; Tishkoff et al., 1996; Campbell and Tishkoff, 2008). One theory suggests that anatomically modern humans may have migrated out of Africa via Ethiopia (Forster and Matsumura, 2005; Reed and Tishkoff, 2006; Campbell and Tishkoff, 2008; Tishkoff et al., 2009). Genetic studies have also provided evidence of a more recent migration into Ethiopia, of Semitic speaking peoples from Arabia (Campbell and Tishkoff, 2008). As a consequence, it is possible that more human genetic/phenotypic variation will be observed in Ethiopians than in any other indigenous group of peoples of similar number living in close geographical proximity. Characterising human genetic variation in Ethiopians may consequently be important for the better design and interpretation of disease association and pharmacogenetic studies and for assisting the reconstruction of human evolutionary history, not only in Ethiopian populations, but also in populations spanning the globe.

Despite the potential importance of Ethiopian population genetics, at the outset of this thesis few studies had investigated the distribution of human genetic variation in Ethiopia, and even then, those that had been undertaken were limited to few populations and small sample sizes. This thesis has contributed to correcting this imbalance by analysing variation in two therapeutically important genes (*CYP1A2* and *TCF7L2*), primarily in a sample set suitable for an initial evaluation of diversity among Ethiopian populations (762 chromosomes from five ethnic groups: Afar, Amhara, Anuak, Maale and Oromo collected from a roughly north east to south west transect across Ethiopia) but variation in *TCF7L2* was also analysed in multiple Ethiopian groups. Data from other populations from publicly available databases were incorporated in the analyses of this thesis to place the Ethiopian data in a worldwide context.

The most striking finding from resequencing *CYP1A2* was that a substantial amount of previously unreported genetic variation was observed in the Ethiopian sample sets. Of particular relevance for healthcare in Ethiopian populations, most individuals may, depending on variation in the promoter region, exhibit normal CYP1A2 profiles. However, several novel rare *CYP1A2* non-synonymous mutations, predicted to cause both conservative and radical changes to protein structure, were identified. The most striking finding from the *TCF7L2* study was that both rs7903146 and rs12255372 alleles, reported to be associated with an increased risk of T2D, are common throughout Ethiopia suggesting that substantial numbers of Ethiopians might be prone to developing T2D. *TCF7L2* haplotype distribution varied among populations suggesting that T2D susceptibility due to the effect of *TCF7L2* may also vary, with most groups likely having a West African *TCF7L2* risk for the disease and some having more of a European *TCF7L2* risk. Subject to associations being confirmed in genotype/phenotype association studies, many *CYP1A2* and *TCF7L2* haplotypes are consequently of important predictive value in the planning and provision of healthcare. These findings are not only of benefit to native

Ethiopians, but are also of increasing importance in the planning of healthcare intervention in the developed world, where growing numbers of individuals with recent Ethiopian descent are living.

Reduction of both intra-population gene diversity and inter-population genetic distance for non-synonymous mutations in comparison to silent mutations (which have no effect on protein structure) in *CYP1A2*, was consistent with the hypothesis that purifying selection has acted at these non-synonymous SNP sites. Further support for this phenomenon comes from the approximate coalescence date boundaries of these non-synonymous mutations being consistent with the following hypothesis: mutations evidenced to be under purifying selection may include variants which drifted to high frequencies in smaller ancestral populations prior to the expansion of anatomically modern humans ~ 100 thousand years ago (Hughes et al., 2003). Since the minor allele frequencies at *CYP1A2* loci evidenced to be under purifying selection are substantially higher than those of genes causative of severe Mendelian diseases, the data suggests that the selective forces acting against these non-synonymous SNPs are mild in comparison to those at SNP sites causative of severe disease. Since mutations associated with complex diseases are expected to be individually only slightly deleterious, as opposed to highly deleterious variants associated with Mendelian diseases, it has been claimed that evidence of mild purifying selection may be used to identify candidate alleles for complex disease-association studies (Hughes et al., 2003). As a consequence, it may be appropriate to include non-synonymous SNPs identified in this thesis in future studies investigating complex diseases which have been linked to *CYP1A2*, e.g. several cancers (Gunes and Dahl, 2008) and cardiovascular disease (Cornelis et al., 2004).

Consistent with Africa being the birthplace of humankind, gene diversity for both *CYP1A2* and *TCF7L2* was always higher in African populations than non-African populations (populations with a recent West African ancestry tended to be the most diverse throughout the analyses, however their sample sizes were often small and this observation should consequently be interpreted with caution). Given that currently most drug testing is undertaken on non-African populations, this thesis has demonstrated that more testing on non European/Asian populations is warranted. With increasing numbers of people having a recent African descent living in Europe and the Americas, their pharmacogenetic profiles should be represented in clinical trials. In addition, there should be close attention paid to them in post-marketing surveillance and greater awareness of variability amongst them. Largely consistent with Y chromosome data, variation in both *CYP1A2* and *TCF7L2* showed an inter-ethnic division between the north Semitic and south non-Semitic speaking groups in Ethiopia. Statistically significant inter-population variation was also observed among Ethiopian populations living in close geographical proximity. As a consequence, the general Ethiopian population should not be treated as a single homogenous group, a finding which undoubtedly has implications for future healthcare policies in Ethiopia.

Serving as further support for the hypothesis that anatomically modern humans expanded out of Africa via Ethiopia, Ethiopian populations evidenced all of the common *CYP1A2* and *TCF7L2* variation observed elsewhere. This thesis demonstrated that CYP1A2 functional variation cannot be predicted in Ethiopians based on known variation, the majority of which is derived from East Asians and Europeans. It may however be possible to use Ethiopians not only as a population for the development of diagnostic tests useful in pharmacogenetic prediction in populations worldwide, but also to ensure that such tests were suitable not only for developed countries.

Consistent with anatomically modern humans having existed in Africa longer than in any other geographic region, much lower LD was observed between loci in each of *CYP1A2* and *TCF7L2* in Africans than in non-Africans. Varying levels of overall low LD between rs7903146 and rs12255372 were also observed among Ethiopian populations with the consequence that rs7903146/rs12255372 haplotypes could effectively discriminate Ethiopian groups in accordance with geography and linguistics, providing insight into inter-population relationships and population histories. The haplotypes even had enough power to differentiate caste-like groups covered by the same self-identifying ethnic label. Ethiopian populations with varying levels of LD may prove particularly informative in T2D genotype/phenotype association studies, and in association studies investigating the causal variant of the association of *TCF7L2* with predisposition to the disease. Likewise, the plethora of variation identified in the *CYP1A2* study should prove of considerable utility in designing association studies investigating pharmacokinetic variation due to the cis promoter region, not least in *CYP1A2* where it has not yet been identified. Similarly, they should prove useful in investigating copy number variation and variability due to mutations in non coding regions e.g. splice sites and initiation sites.

I have been very fortunate in the timing of my PhD research. Ethiopia has increasingly opened up its doors to population genetic research. I owe a special thanks to Dr Ayele Tarekegn, without whom this research would not be possible. As a postdoctoral research fellow affiliated with Addis Ababa University, he has spent the past 12 years travelling far and wide in Ethiopia collecting buccal DNA and sociological background data from over 8000 Ethiopians from multiple ethnic groups. From personal fieldwork experience in Ethiopia, this is certainly not an easy task. The studies included in this thesis also come at a time when the UK government is reporting steps which have to be taken to ensure that the NHS is able to exploit the benefits of genetic research (http://www.parliament.uk/hlscience/). I hope future reports push forward the potential healthcare benefits of approaches discussed in this thesis.

6.1 Future work

The outcomes of the studies included in this thesis suggest a number of potential projects that are immediately apparent, some of which are listed below:

- Sequence more CYP genes in the Ethiopian ascertainment populations (refer to discussion of chapter 4 for details of genes).

- Since the Ethiopian data could only be tentatively interpreted within the context of West Africa, due to small Yoruba sample sizes, extend the *CYP1A2* study to a number of West African populations with larger sample sizes.

- Develop a diagnostic test for predicting CYP1A2 functional variation among Ethiopian populations (refer to chapter 4 discussion for further details).

- Confirm whether the rs7903146/rs12255372 TG haplotype is significantly associated with an increased risk of T2D in Afro-Caribbeans by repeating the Humphries et al. (2006) study with larger sample sizes. If the association is confirmed, establish whether typing additional SNPs to rs7903146 and rs12255372 increases the association of *TCF7L2* with T2D in Afro-Caribbeans, using Afro-Caribbeans without T2D as a control.

- Characterise human genetic variation in Ethiopia at a much finer scale by typing the Ethiopian ascertainment populations with the million SNP chip – a project which is now being planned.

- Whole genome sequencing of representative donors from the groups analysed in this thesis – a project which is also in the course of being planned.

Appendix 1

From genotype to haplotype

A haplotype is *the combination of allelic states of a set of polymorphic markers lying on the same DNA molecule e.g. a chromosome or region of a chromosome* (Jobling et al., 2004). Haplotypes are immediately identified from haploid loci (mtDNA, Y chromosome and the X chromosome in males) but are less readily obtained from diploid autosomal loci. The traditional method for determining haplotypes is by pedigree analysis, but this may be uninformative depending on allele frequencies within the pedigree. Haplotypes can be obtained from genotype data by physical separation of one allele from the other, but this approach can be expensive and laborious (Jobling et al., 2004). As a consequence of this, statisticians have developed, and continue to do so, algorithms for haplotype reconstruction. Common methods include Clark's algorithm, the expectation-maximization (EM) algorithm, the ELB algorithm and a Bayesian approach implemented in PHASE. A newer model using the EM algorithm is implemented in fastPHASE.

Clark's algorithm

One of the earliest algorithms for haplotype reconstruction, from genotype data, was described in Clark (1990). The approach is based on the principle of maximum parsimony as it attempts to minimise the total number of steps required to resolve a haplotype. The rationale for Clark's algorithm is that an ambiguous pair of haplotypes is likely to contain at least one common haplotype. The algorithm starts by identifying all unambiguous haplotypes (from all homozygotes and single-site heterozygotes) and considering them as resolved. The next step determines whether any of the resolved haplotypes could be one of the underlying haplotypes in the remaining unphased genotypes. If this is the case, the haplotype is assumed to be correct and the complementary haplotype is added to the set of resolved haplotypes. The algorithm continues to analyse the data, increasing the pool of resolved haplotypes, until all genotypes are resolved or no further genotypes can be resolved in this way.

When manually applying Clark's algorithm to the *CYP1A2* genotype data, the following extensions were applied in the Clark-ancestral approach:-

1. For each individual, if the unphased genotype data can be divided into two previously resolved haplotypes, then this is the pair of haplotypes to assign to that individual.

2. When none of the possible haplotypes are in the resolved set, refer to the chimpanzee's combination of alleles for this locus. If possible, assign the chimpanzee's haplotype, plus the complementary haplotype, to this individual and add them to the resolved haplotype set.

3. When one of the possible haplotypes is that of the chimpanzee (which has already been added to the resolved haplotype set), always adopt this option, unless the unphased genotype data can be divided into two previously resolved haplotypes.

4. When one of the possible haplotypes can be a number of different haplotypes from the resolved set, and the status of the chimpanzee is unknown or not a possibility, assign the individual to the most common haplotype in the sample population, plus the complementary haplotype.

5. When none of the above apply and none of the possible haplotypes are in the resolved set, assign this individual as having two orphan haplotypes.

Another extension of Clark's algorithm (Clark-common) was also manually applied to the *CYP1A2* data. It is a certain resolution approach and underestimates diversity by not resolving low frequency polymorphisms into haplotypes. Unlike the Clark-ancestral approach described above, the Clark-common approach prioritises the common haplotype in the population over the ancestral haplotype. Thus when one of the possible haplotypes can be a number of different haplotypes from the resolved set, the individual is assigned to the most common haplotype in the sample population (whether it is the ancestral haplotype or not), plus the complementary haplotype. The Clark-ancestral and Clark-common approaches were applied to the *CYP1A2* data in a segmented fashion. Haplotypes were deduced separately for each exon and intron, and for the 5' upstream region of the gene and the 3' UTR, and were then linked together to produce haplotypes for the entire *CYP1A2* gene.

The advantages of Clark's algorithm are that it is a relatively simple procedure which can handle large numbers of loci when haplotype diversity is limited in the population. The algorithm also performs well for short stretches of DNA and longer regions in complete LD. Along with the possibility of "orphan" haplotypes remaining undeduced, disadvantages include the algorithm not starting (at least not without extra information or adhoc intervention) when there are no homozygotes or single-site heterozygotes in the population, the algorithm not giving unique solutions due to the phased haplotypes being dependant on the order of the unphased genotypes, missing data not being accounted for, and its performance is relatively sensitive to the extent of deviation from Hardy-Weinberg equilibrium, although the approach does not explicitly assume Hardy-Weinberg equilibrium (Niu et al., 2002). In addition, Clark's algorithm cannot assess the uncertainty associated with each phase call.

Expectation-Maximization (EM) algorithm

The maximum-likelihood approach, implemented through the EM algorithm (Excoffier and Slatkin, 1995), is an iterative procedure which attempts to find the set of unknown population haplotype frequencies that maximises the likelihood of observing the known sample genotype

frequencies, under the assumption of Hardy-Weinberg equilibrium. The estimate found by the EM algorithm can depend on the starting point, so in practice a number of different starting points are used. The first starting point has, as an example, been computed by finding all the possible haplotypes which could occur in the sample population given the genotypes, and setting each of them to an equal frequency (Excoffier and Slatkin, 1995). Once the matrices of estimated haplotype frequencies are obtained, the haplotypes are then reconstructed for each individual generally by choosing the most probable haplotype assignment, given the estimated haplotype frequencies and the genotype data.

Although the EM method is more computationally intensive than Clark's algorithm, it has the advantage of estimating haplotype frequencies from a solid statistical theory. Other advantages of the EM approach include its performance not being too sensitive to departures from Hardy-Weinberg equilibrium, despite it making a patent assumption of Hardy-Weinberg equilibrium (Niu et al., 2002). Disadvantages include the standard EM algorithm not being able to handle a large number of loci due to the need to store haplotype frequency estimates for every possible haplotype in the sample.

ELB algorithm

The ELB algorithm (Excoffier et al., 2003), implemented in Arelquin software (Schneider et al., 2000), reconstructs the gametic phase of multiple genotypes. A window of neighbouring loci is used to create phase updates, and the size of the window changes according to the local level of linkage disequilibrium; the higher the linkage disequilibrium between pairs of loci, the larger the window and vice versa. The ELB algorithm begins by arbitrarily assigning phase to each individual in the population and associating each heterozygous locus with a window containing the locus and neighboring loci. At each iteration of the algorithm, an individual is randomly chosen, and one by one, its heterozygous loci are randomly visited. For each locus, the window is updated by accepting or rejecting, (a) an extra locus at one end of the window and (b) the withdrawal of a locus at the opposite end. The locus being visited always remains inside the window and at least one other heterozygous locus is always included in each window. The two proposed updates are made successively so that the window can either increase by one locus, decrease by one locus, increase by one locus in either direction (if both changes are accepted) or remain unchanged (if both changes are rejected). Finally, phase updates for each locus being visited are made based on haplotypes observed in corresponding windows of other individuals in the sample population.

The key advantage of the ELB algorithm is that it is suited to problems involving multiple loci and/or large genomic regions, including those with recombination. The approach also evidenced consistently better accuracy on a local scale than other methods, such as PHASE, while its accuracy over larger genomic regions was close to the best rates (Excoffier et al., 2003). Missing data can be handled and haplotypes are reconstructed quickly when small

numbers of loci are considered but slowly when large datasets are used. Probability estimates for each haplotype pair per sample are also calculated.

PHASE

The program PHASE (Stephens et al., 2001) implements a Bayesian statistical method for reconstructing haplotypes from population genotype data. The algorithm begins by identifying all unambiguous haplotypes from homozygotes and single-site heterozygotes in the sample. A posterior distribution (*the conditional distribution of the unobserved haplotypes given the observed genotype data* (Stephens and Donnelly, 2003)) is then constructed. This is based on a prior that approximates the coalescent (Stephens and Donnelly, 2003), the rationale being that the difference in genetic sequence between the progenitor and mutant offspring will be small, often only by a single base change. The prior will put substantial weight on the possibility of dividing unphased genotypes into two haplotypes, one or both of which are already known in the sample. When unphased genotypes cannot be divided in this manner, the approximate coalescent prior will put substantial weight on the possibility of both haplotypes being similar to known haplotypes. Once the posterior distribution has been constructed, the Markov-chain Monte Carlo method makes an initial guess, from this distribution, at all haplotype pairs in the sample, except one. This individual is then assigned a pair of haplotypes from the posterior distribution on the assumption that all other haplotype pairs in the sample are correct (on occasions where none of the haplotypes from the posterior distribution are complementary to the unphased genotype data, the algorithm randomly sprinkles mutations onto the haplotypes which it chooses). The algorithm then estimates the probability of this reconstruction being correct given the genotype of the individual. Sufficient repetition of this process creates a distribution, per individual, of the probability of the reconstructed haplotypes given the genotypes. At the end of the algorithm, individuals are assigned haplotypes which are the most probable given the genotype data. The algorithm initially estimates haplotype frequencies in short blocks of consecutive SNPs. Adjacent blocks are then combined and haplotype estimates are obtained for the entire region under consideration.

The key advantage of the algorithm is that it integrates the coalescence theory into its prior, and performs well in simulations based on a coalescent model. The algorithm can also impute missing data and experiments on both real and simulated data show that it can outperform other methods when reconstructing haplotypes (Stephens et al., 2001). It is not however clear whether the algorithm outperforms other methods for admixed or rapidly expanding populations when the coalescent theory is not supported. PHASE and EM based approaches produce similar results in in such cases (Zhang et al., 2001; Xu et al., 2002). Stephens and Donnelly (2003) suggest increasing PHASE run times to increase the accuracy of haplotypes in these circumstances. The main disadvantage of the algorithm is its lack of speed. The algorithm works slowly, as it adopts a “piece by piece” strategy when updating new haplotypes which resemble existing haplotypes.

fastPHASE

The programme fastPHASE implements methods described in Scheet and Stephens (2006). It is a statistical model for haplotype reconstruction from population genotype data on a large scale. The model is based on the theory that haplotypes tend to cluster into groups of similar haplotypes in a population over short regions (~ 3 kb in humans), with each cluster representing a common haplotype across the entire region under consideration. This clustering is local in nature due to recombination, such that as one moves along the chromosome, similar haplotypes will vary. fastPHASE allows cluster memberships of observed haplotypes to continuously vary along the chromosome via a hidden Markov model. Each observed haplotype ends up as a mosaic of a limited number of common haplotypes in the population. Unlike block based cluster models, which separate the chromosome into segments of high linkage disequilibrium only allowing cluster memberships to vary across block boundaries, fastPHASE is flexible and captures more complex patterns of linkage disequilibrium while continuing to pick up on any block based patterns that may be present. As examples, the model can capture the immediate break down of linkage disequilibrium across recombination hotspots and the more moderate decline of linkage disequilibrium with distance.

The algorithm starts by assigning each observed sample to *n* clusters of origin by weighted probability (eg. sample 1 is more likely to come from cluster x, because cluster x is high in frequency) and given the genotype data. The EM algorithm is then implemented to construct a matrix of the frequencies of each SNP in each cluster. The probability of the observed haplotype is estimated given the matrix and frequency of clusters. Each haplotype is then assigned a probability of which cluster(s) it belongs to. The algorithm then estimates the probability of each SNP, in each haplotype, originating from each cluster. At this point, the hidden Markov model is used to take neighbouring SNPs into account as well as the distance between markers. The algorithm then calculates probabilities of specific genotypes originating from specific clusters. For each individual, the EM algorithm is then employed to create a distribution of haplotype probabilities given the matrices, frequency of clusters and genotype data. At the end of the process, the most probable haplotypes are assigned to each individual.

The advantages of the model are that it is flexible, fast, computationally convenient and can handle large amounts of data (thousands of individuals with hundreds of thousands of SNPs). It can impute missing genotypes as accurately, or more so, than existing methods such as PHASE. Disadvantages of the model include its haplotype estimates being slightly less accurate than those from PHASE, it not producing recombination rate estimates, and it only being a purely predictive model. The algorithm only estimates haplotypes. It does not relate genetic data to underlying models of demography or evolution. In addition, despite imputing missing genotypes, the model is only suited to imputation of genotypes at SNPs where many individuals have been genotyped (Scheet and Stephens, 2006).

Supplementary Data

Supplementary figure S1 Population pairwise genetic distances (pink) and p values (upper triangle) for *CYP1A2* haplotypes. P values below the 5 % significance threshold are highlighted in grey.

	CYP1A2 (entire gene) haplotypes									
	Afar	Amhara	Anuak	Maale	Oromo	African American	Yoruba	European	Hispanic	East Asian
Afar	-0.01	0.34	< 0.01	< 0.01	0.25	0.14	0.02	< 0.01	0.04	0.05
Amhara	0.00	-0.01	< 0.01	< 0.01	0.07	0.04	< 0.01	0.01	0.09	0.02
Anuak	0.08	0.12	-0.01	< 0.01	< 0.01	0.08	0.43	< 0.01	< 0.01	< 0.01
Maale	0.03	0.06	0.02	-0.01	< 0.01	0.16	0.12	< 0.01	< 0.01	< 0.01
Oromo	0.00	0.01	0.08	0.02	-0.01	0.35	0.11	< 0.01	0.09	0.16
African American	0.02	0.05	0.03	0.01	0.00	-0.08	0.63	< 0.01	0.02	0.73
Yoruba	0.07	0.12	0.00	0.03	0.04	-0.02	-0.11	< 0.01	< 0.01	0.29
European	0.11	0.08	0.34	0.21	0.13	0.27	0.39	-0.04	0.27	< 0.01
Hispanic	0.05	0.04	0.27	0.15	0.05	0.13	0.23	0.01	-0.08	0.05
East Asian	0.03	0.06	0.08	0.04	0.01	-0.03	0.03	0.26	0.11	-0.04

	CYP1A2 cds (non-synonymous variants) haplotypes									
	Afar	Amhara	Anuak	Maale	Oromo	African American	Yoruba	European	Hispanic	East Asian
Afar	-0.01	0.04	0.11	0.05	0.79	0.75	0.32	0.24	0.27	0.26
Amhara	0.02	-0.01	0.01	< 0.01	0.12	0.28	0.08	0.99	0.99	0.99
Anuak	0.01	0.05	-0.01	0.17	0.15	0.77	0.77	0.03	0.13	0.06
Maale	0.01	0.05	0.00	-0.01	0.03	0.45	0.58	0.05	0.10	0.05
Oromo	0.00	0.01	0.01	0.01	-0.01	0.79	0.21	0.25	0.41	0.44
African American	-0.02	0.03	-0.01	-0.01	-0.02	-0.05	0.77	0.17	0.50	0.51
Yoruba	0.01	0.15	-0.03	-0.02	0.02	-0.03	-0.09	0.07	0.14	0.10
European	0.01	-0.01	0.04	0.04	0.01	0.04	0.15	-0.03	0.99	0.99
Hispanic	0.01	-0.02	0.03	0.03	0.00	0.02	0.11	0.00	-0.05	0.99
East Asian	0.00	-0.01	0.03	0.03	0.00	0.01	0.08	-0.01	-0.01	-0.03

Supplementary figure S2 Population pairwise genetic distances (Fst) (pink) and p values (upper triangle) for *TCF7L2* haplotypes. P values below the 5 % significance threshold are highlighted in grey.

	Afro-Caribbeans with T2D	Yoruba	European	Chinese	Japanese
Afro-Caribbeans with T2D	-0.006	0.796	p < 0.001	p < 0.001	p < 0.001
Yoruba	-0.004	-0.008	p < 0.001	p < 0.001	p < 0.001
European	0.136	0.145	-0.008	p < 0.001	p < 0.001
Chinese	0.312	0.334	0.163	-0.011	0.707
Japanese	0.299	0.319	0.148	-0.007	-0.011

Supplementary figure S3 Population pairwise genetic distances (Fst) (pink) and p values (upper triangle) for *TCF7L2* haplotypes. P values below the 5 % significance threshold are highlighted in grey.

a) Based on seven SNP haplotype

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar	-1.01	0.14	p<0.01	p<0.01	0.08	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01
Amhara	0.01	-1.01	p<0.01	p<0.01	0.39	p<0.01	p<0.01	0.03	p<0.01	p<0.01
Anuak	0.04	0.07	-1.01	0.05	p<0.01	0.02	0.16	p<0.01	p<0.01	p<0.01
Maale	0.02	0.04	0.01	-1.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01
Oromo	0.01	0.00	0.05	0.03	-1.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01
Afro-Caribbeans	0.04	0.06	0.01	0.02	0.04	-1.01	0.82	p<0.01	p<0.01	p<0.01
Yoruba	0.04	0.07	0.00	0.03	0.05	0.00	-1.01	p<0.01	p<0.01	p<0.01
European	0.03	0.02	0.14	0.10	0.04	0.14	0.15	-1.01	p<0.01	p<0.01
Chinese	0.19	0.21	0.30	0.25	0.25	0.31	0.33	0.16	-1.01	0.73
Japanese	0.18	0.19	0.28	0.24	0.23	0.30	0.32	0.15	-0.01	-1.01

b) Based on two SNP haplotype

	Afar	Amhara	Anuak	Maale	Oromo	Afro-Caribbeans	Yoruba	European	Chinese	Japanese
Afar	-1.01	0.19	0.11	0.04	0.07	0.03	0.02	p<0.01	p<0.01	p<0.01
Amhara	0.01	-1.01	0.03	p<0.01	0.81	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01
Anuak	0.01	0.02	-1.01	0.34	p<0.01	0.23	0.29	p<0.01	p<0.01	p<0.01
Maale	0.02	0.02	0.00	-1.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01
Oromo	0.01	0.00	0.02	0.03	-1.01	p<0.01	p<0.01	p<0.01	p<0.01	p<0.01
Afro-Caribbeans	0.02	0.03	0.00	0.02	0.03	-1.01	0.76	p<0.01	p<0.01	p<0.01
Yoruba	0.02	0.04	0.00	0.02	0.04	-0.01	-1.01	p<0.01	p<0.01	p<0.01
European	0.05	0.04	0.10	0.10	0.05	0.12	0.13	-1.01	p<0.01	p<0.01
Chinese	0.19	0.22	0.25	0.24	0.24	0.28	0.29	0.16	-1.01	0.76
Japanese	0.17	0.20	0.24	0.23	0.23	0.27	0.28	0.14	-0.01	-1.01

Supplementary figure S4 Population pairwise genetic distances (Fst) for rs7903146/rs12255372 haplotypes. Significant distances are highlighted in pink ($p < 0.05$). Afro-Caribbeans have T2D. Basketo are also known as Masketo and Bume are also known as Nyangagtom.

	Ari	Ari Cultivator	Afar	Agew, Eastern	Alaba	Amhara	Anuak	Ari Potter	Ari Smith	Bena	Basketo	Bume	Bench	Burji	Busa	Dizi	Dime	Dirasha	Dasenech	Dawuro	Dorze	Gobeze	Gedeo	Gofa	Ganjule	Gamo	Genta	Gurage	Gewada	Hadiya	Hamer	Kefa	Kembata	Konta	Kore	Konso	Manja	Mejenger	Mashile	Maale	Nuer	Oromo	Shekecho	Sidama	Sheko	Somali	Tigrayan	Wolayta	Yem	Zayse	Afro-Caribbean	Yoruba	European	Chinese	Japanese
Ari	-0.004																																																						
Ari Cultivator	0.001	-0.002																																																					
Afar	0.078	0.116	-0.004																																																			
Agew, Eastern	0.086	0.123	0.000	-0.002																																																		
Alaba	0.053	0.086	0.003	0.002	-0.004																																																	
Amhara	0.076	0.110	0.004	0.000	0.000	-0.001																																																
Anuak	0.070	0.106	-0.003	0.005	0.005	0.010	-0.005																																														
Ari Potter	0.004	0.021	0.040	0.046	0.021	0.040	0.033	-0.004																																													
Ari Smith	0.040	0.067	0.034	0.047	0.030	0.047	0.020	0.016	-0.010																																											
Bena	-0.016	-0.009	0.062	0.067	0.035	0.056	0.059	-0.007	0.048	-0.034																																										
Basketo	-0.002	0.010	0.052	0.059	0.032	0.052	0.045	-0.002	0.027	-0.014	-0.004																																									
Bume	0.044	0.077	-0.002	0.003	0.000	0.004	0.001	0.020	0.040	0.018	0.023	-0.021																																									
Bench	-0.002	0.007	0.065	0.073	0.043	0.066	0.055	0.000	0.025	-0.011	-0.003	0.038	-0.004																																								
Burji	0.006	0.023	0.041	0.051	0.027	0.047	0.032	-0.002	0.009	-0.001	0.000	0.022	0.000	-0.004																																							
Busa	0.007	0.026	0.035	0.043	0.019	0.038	0.027	-0.003	0.010	-0.002	-0.001	0.017	0.001	-0.003	-0.004																																							
Dizi	0.009	0.028	0.032	0.041	0.018	0.036	0.025	-0.003	0.009	-0.001	0.000	0.015	0.002	-0.003	-0.004	-0.004																																							
Dime	0.001	0.018	0.032	0.038	0.014	0.031	0.027	-0.007	0.019	-0.014	-0.006	0.007	-0.002	-0.004	-0.006	-0.006	-0.012																																							
Dirasha	0.013	0.034	0.028	0.035	0.013	0.030	0.020	-0.002	0.007	0.004	0.003	0.014	0.005	-0.002	-0.002	-0.004	-0.004	-0.005	-0.005																																							
Dasenech	0.123	0.165	0.009	0.015	0.023	0.023	0.006	0.072	0.037	0.130	0.092	0.041	0.103	0.071	0.064	0.061	0.073	0.052	-0.019																																							
Dawuro	0.046	0.076	0.004	0.005	-0.002	0.003	0.005	0.018	0.027	0.027	0.026	-0.004	0.037	0.022	0.016	0.015	0.010	0.011	0.027	-0.002																																						
Dorze	0.002	0.016	0.042	0.048	0.023	0.041	0.037	-0.003	0.026	-0.014	-0.004	0.014	0.000	0.000	-0.001	-0.001	-0.008	0.001	0.082	0.018	-0.005																																						
Gobeze	0.024	0.050	0.014	0.018	0.003	0.015	0.010	0.003	0.011	0.011	0.010	0.003	0.016	0.005	0.001	0.001	-0.002	-0.002	0.037	0.002	0.005	-0.004																																						
Gedeo	-0.001	0.012	0.051	0.060	0.033	0.053	0.044	-0.002	0.023	-0.012	-0.004	0.023	-0.003	-0.002	-0.001	0.000	-0.006	0.002	0.090	0.027	-0.003	0.010	-0.004																																						
Gofa	0.021	0.045	0.017	0.021	0.004	0.017	0.014	0.001	0.013	0.007	0.007	0.005	0.013	0.004	0.000	0.000	-0.004	-0.002	0.043	0.003	0.003	-0.004	0.008	-0.004																																						
Ganjule	0.011	0.032	0.031	0.038	0.016	0.033	0.023	-0.003	0.006	0.003	0.002	0.016	0.004	-0.003	-0.004	-0.004	-0.005	-0.005	0.056	0.013	0.001	-0.001	0.001	-0.001	-0.005																																						
Gamo	0.044	0.075	0.003	0.007	-0.001	0.007	0.001	0.015	0.017	0.030	0.025	-0.002	0.034	0.017	0.012	0.011	0.009	0.007	0.021	-0.001	0.018	0.000	0.025	0.001	0.009	-0.002																																						
Genta	0.012	0.032	0.024	0.029	0.010	0.024	0.020	-0.002	0.018	-0.004	0.001	0.004	0.007	0.001	-0.002	-0.002	-0.007	-0.002	0.057	0.007	-0.002	-0.002	0.002	-0.003	-0.002	0.006	-0.004																																					
Gurage	0.069	0.105	0.000	-0.001	-0.002	-0.001	0.002	0.033	0.034	0.052	0.045	0.001	0.057	0.037	0.029	0.028	0.025	0.022	0.015	0.000	0.035	0.009	0.046	0.011	0.025	0.001	0.018	-0.003																																					
Gewada	0.008	0.027	0.031	0.040	0.018	0.035	0.024	-0.003	0.012	-0.004	-0.001	0.011	0.002	-0.003	-0.004	-0.004	-0.007	-0.003	0.062	0.014	-0.002	0.001	-0.001	0.000	-0.003	0.010	-0.003	0.027	-0.004																																					
Hadiya	0.041	0.072	0.002	0.007	-0.001	0.007	0.001	0.014	0.017	0.026	0.022	-0.005	0.031	0.015	0.011	0.009	0.007	0.006	0.024	-0.001	0.015	-0.001	0.022	0.001	0.008	-0.003	0.004	0.001	0.009	-0.004																																					
Hamer	0.002	0.023	0.019	0.028	0.008	0.025	0.013	-0.011	0.003	-0.011	-0.008	-0.001	-0.005	-0.011	-0.012	-0.012	-0.015	-0.012	0.056	0.003	-0.010	-0.009	-0.009	-0.009	-0.012	0.000	-0.011	0.016	-0.013	-0.002	-0.022																																					
Kefa	0.040	0.068	0.017	0.013	0.002	0.006	0.021	0.018	0.046	0.016	0.024	0.000	0.036	0.028	0.019	0.019	0.008	0.016	0.052	0.002	0.015	0.006	0.026	0.006	0.018	0.007	0.007	0.007	0.017	0.006	0.008	-0.004																																					
Kembata	0.046	0.078	0.001	0.005	-0.002	0.006	-0.001	0.016	0.017	0.032	0.026	-0.003	0.035	0.018	0.013	0.011	0.010	0.008	0.019	-0.001	0.019	0.000	0.026	0.002	0.010	-0.003	0.006	0.000	0.011	-0.004	0.000	0.007	-0.004																																					
Konta	0.011	0.028	0.036	0.038	0.016	0.030	0.036	0.004	0.041	-0.011	0.003	0.007	0.012	0.011	0.007	0.007	-0.004	0.008	0.080	0.013	-0.001	0.007	0.005	0.004	0.009	0.016	0.000	0.027	0.005	0.014	-0.003	0.004	0.017	-0.005																																					
Kore	0.010	0.029	0.032	0.040	0.017	0.035	0.025	-0.003	0.008	0.001	0.001	0.017	0.003	-0.003	-0.004	-0.004	-0.006	-0.004	0.059	0.014	0.000	0.000	0.001	-0.001	-0.004	0.010	-0.002	0.027	-0.003	0.009	-0.012	0.018	0.011	0.008	-0.004																																					
Konso	0.038	0.068	0.005	0.011	0.000	0.010	0.002	0.011	0.011	0.026	0.020	0.000	0.027	0.011	0.007	0.006	0.006	0.003	0.023	0.000	0.014	-0.002	0.019	-0.001	0.004	-0.003	0.003	0.003	0.006	-0.003	-0.004	0.009	-0.004	0.015	0.006	-0.004																																					
Manja	0.003	0.018	0.042	0.049	0.024	0.043	0.035	-0.003	0.020	-0.010	-0.003	0.017	0.000	-0.001	-0.002	-0.002	-0.007	0.000	0.077	0.020	-0.003	0.005	-0.003	0.003	-0.001	0.017	-0.001	0.035	-0.003	0.015	-0.010	0.019	0.018	0.003	-0.001	0.013	-0.003																																					
Mejenger	0.037	0.066	0.012	0.023	0.011	0.024	0.004	0.011	0.001	0.031	0.020	0.008	0.024	0.007	0.006	0.005	0.008	0.003	0.025	0.009	0.016	0.001	0.018	0.003	0.003	0.003	0.007	0.014	0.005	0.002	-0.005	0.023	0.002	0.024	0.005	-0.001	0.013	-0.004																																					
Mashile	0.044	0.074	0.009	0.016	0.005	0.017	0.003	0.014	0.004	0.037	0.025	0.010	0.030	0.013	0.009	0.008	0.011	0.004	0.018	0.006	0.020	0.000	0.024	0.002	0.005	0.000	0.008	0.008	0.009	0.000	-0.001	0.019	-0.001	0.025	0.007	-0.002	0.017	-0.002	-0.004																																					
Maale	0.017	0.040	0.021	0.029	0.012	0.027	0.014	0.000	0.007	0.006	0.005	0.006	0.009	-0.001	-0.002	-0.003	-0.004	-0.003	0.046	0.008	0.002	-0.002	0.004	-0.002	-0.003	0.004	-0.002	0.019	-0.003	0.003	-0.012	0.015	0.004	0.008	-0.002	0.001	0.001	0.000	0.003	-0.004																																					
Nuer	0.080	0.117	0.003	0.011	0.010	0.016	-0.002	0.039	0.015	0.077	0.054	0.016	0.062	0.037	0.032	0.030	0.037	0.024	-0.001	0.012	0.047	0.014	0.052	0.019	0.026	0.006	0.028	0.007	0.031	0.007	0.020	0.032	0.004	0.049	0.029	0.005	0.043	0.005	0.002	0.019	-0.004																																					
Oromo	0.083	0.120	0.003	-0.001	0.002	-0.002	0.011	0.046	0.056	0.060	0.058	0.002	0.074	0.054	0.044	0.042	0.036	0.037	0.025	0.005	0.045	0.019	0.059	0.021	0.040	0.009	0.028	0.000	0.041	0.009	0.029	0.007	0.008	0.032	0.041	0.013	0.048	0.029	0.021	0.032	0.019	-0.003																																					
Shekecho	0.024	0.047	0.019	0.020	0.005	0.015	0.020	0.008	0.036	0.001	0.010	-0.004	0.021	0.014	0.009	0.008	-0.001	0.008	0.058	0.003	0.004	0.002	0.012	0.002	0.009	0.006	0.000	0.013	0.006	0.004	-0.003	-0.001	0.007	-0.002	0.009	0.006	0.007	0.016	0.015	0.006	0.032	0.016	-0.004																																					
Sidama	0.032	0.059	0.010	0.011	-0.001	0.007	0.010	0.008	0.023	0.014	0.015	-0.001	0.024	0.013	0.007	0.007	0.001	0.004	0.037	-0.001	0.009	-0.002	0.017	-0.002	0.006	-0.001	0.000	0.004	0.006	-0.001	-0.003	0.000	-0.001	0.005	0.006	-0.001	0.010	0.007	0.004	0.003	0.017	0.010	-0.001	-0.004																																					
Sheko	-0.003	-0.002	0.097	0.104	0.069	0.093	0.088	0.012	0.055	-0.014	0.003	0.058	0.002	0.014	0.016	0.018	0.009	0.023	0.147	0.060	0.008	0.037	0.004	0.032	0.021	0.060	0.021	0.087	0.016	0.056	0.011	0.053	0.062	0.018	0.019	0.053	0.009	0.051	0.060	0.028	0.100	0.101	0.034	0.045	-0.004																																					
Somali	0.050	0.083	0.002	0.008	0.001	0.009	-0.002	0.018	0.011	0.039	0.029	0.003	0.037	0.018	0.013	0.012	0.013	0.008	0.013	0.001	0.022	0.001	0.028	0.003	0.009	-0.002	0.009	0.002	0.012	-0.003	0.002	0.013	-0.004	0.023	0.011	-0.003	0.021	-0.001	-0.003	0.005	0.000	0.012	0.012	0.002	0.067	-0.005																																					
Tigrayan	0.095	0.135	-0.003	-0.003	0.003	0.001	0.001	0.051	0.044	0.080	0.066	0.008	0.081	0.055	0.046	0.044	0.044	0.037	0.003	0.007	0.054	0.020	0.066	0.024	0.040	0.006	0.034	-0.002	0.043	0.007	0.032	0.019	0.004	0.046	0.043	0.010	0.054	0.021	0.013	0.031	0.004	0.000	0.026	0.014	0.115	0.005	-0.007																																					
Wolayta	0.014	0.035	0.022	0.027	0.008	0.022	0.018	-0.001	0.016	-0.001	0.002	0.004	0.008	0.001	-0.002	-0.002	-0.007	-0.003	0.053	0.006	-0.001	-0.003	0.003	-0.004	-0.002	0.004	-0.004	0.016	-0.002	0.003	-0.011	0.007	0.005	0.001	-0.002	0.002	0.000	0.005	0.006	-0.002	0.025	0.026	0.000	-0.001	0.024	0.007	0.031	-0.004																																					
Yem	0.007	0.025	0.032	0.037	0.015	0.031	0.027	-0.003	0.019	-0.007	-0.001	0.011	0.003	-0.001	-0.003	-0.003	-0.008	-0.002	0.066	0.012	-0.004	0.000	-0.001	-0.002	-0.002	0.010	-0.004	0.025	-0.003	0.009	-0.011	0.011	0.011	0.000	-0.003	0.007	-0.003	0.010	0.011	-0.001	0.035	0.036	0.002	0.003	0.015	0.014	0.042	-0.004	-0.005																																					
Zayse	0.002	0.018	0.040	0.046	0.021	0.040	0.034	-0.004	0.021	-0.011	-0.003	0.016	0.000	-0.001	-0.003	-0.002	-0.008	-0.001	0.076	0.017	-0.004	0.003	-0.003	0.001	-0.001	0.016	-0.002	0.033	-0.003	0.014	-0.011	0.016	0.017	0.001	-0.002	0.012	-0.004	0.013	0.016	0.001	0.042	0.045	0.005	0.008	0.009	0.019	0.052	-0.002	-0.004	-0.005																																					
Afro-Caribbean	0.051	0.078	0.026	0.035	0.030	0.038	0.028	0.037	0.060	0.024	0.034	-0.007	0.049	0.037	0.035	0.032	0.023	0.035	0.075	0.023	0.028	0.028	0.034	0.029	0.037	0.025	0.023	0.034	0.027	0.020	0.014	0.025	0.024	0.022	0.037	0.026	0.032	0.030	0.039	0.024	0.048	0.034	0.015	0.025	0.061	0.031	0.042	0.024	0.029	0.032	-0.006																																					
Yoruba	0.051	0.078	0.026	0.035	0.030	0.038	0.028	0.037	0.060	0.024	0.034	-0.007	0.049	0.037	0.035	0.032	0.023	0.035	0.075	0.023	0.028	0.028	0.034	0.029	0.037	0.025	0.023	0.034	0.027	0.020	0.014	0.025	0.024	0.022	0.037	0.026	0.032	0.030	0.039	0.024	0.048	0.034	0.015	0.025	0.061	0.031	0.042	0.024	0.029	0.032	-0.006	-0.008																																					
European	0.199	0.241	0.039	0.031	0.053	0.037	0.049	0.138	0.124	0.211	0.162	0.086	0.180	0.145	0.132	0.128	0.142	0.118	0.016	0.060	0.145	0.089	0.162	0.096	0.123	0.062	0.114	0.039	0.129	0.067	0.133	0.080	0.061	0.130	0.127	0.072	0.141	0.090	0.073	0.110	0.046	0.036	0.099	0.078	0.225	0.060	0.022	0.109	0.127	0.141	0.124	0.124	-0.008																																					
Chinese	0.340	0.359	0.170	0.161	0.205	0.170	0.168	0.272	0.261	0.553	0.303	0.349	0.311	0.270	0.261	0.255	0.353	0.253	0.167	0.183	0.295	0.227	0.295	0.239	0.257	0.183	0.258	0.178	0.262	0.199	0.392	0.247	0.196	0.293	0.260	0.202	0.266	0.202	0.187	0.236	0.150	0.190	0.253	0.225	0.368	0.189	0.182	0.253	0.272	0.285	0.279	0.279	0.158	-0.011																																					
Japanese	0.334	0.355	0.160	0.152	0.195	0.162	0.159	0.266	0.252	0.524	0.296	0.323	0.305	0.264	0.255	0.248	0.338	0.246	0.146	0.176	0.287	0.219	0.288	0.231	0.250	0.175	0.250	0.169	0.255	0.191	0.369	0.237	0.187	0.284	0.253	0.193	0.260	0.195	0.180	0.229	0.142	0.180	0.244	0.216	0.362	0.180	0.168	0.245	0.265	0.278	0.266	0.266	0.138	-0.007	-0.011	

Supplementary figure S5 Pairwise geographic distances (km) among multiple Ethiopian populations. Basketo are also known as Masketo and Bume are also known as Nyangagtom.

	Ari	Ari Cultivator	Afar	Agew, Eastern	Alaba	Amhara	Anuak	Ari Potter	Ari Smith	Bena	Basketo	Bume	Bench	Burji	Busa	Dizi	Dime	Dirasha	Dasenech	Dawuro	Dorze	Gobeze	Gedeo	Gofa	Ganjule	Gamo	Genta	Gurage	Gewada	Hadiya	Hamer	Kefa	Kembata	Konta	Kore	Konso	Manja	Mejenger	Mashile	Maale	Nuer	Oromo	Shekecho	Sidama	Sheko	Somali	Tigrayan	Wolayta	Yem
Ari Cultivator	26.38																																															
Afar	833.86	826.95																																														
Agew, Eastern	660.20	642.89	373.61																																													
Alaba	236.63	231.89	597.55	465.69																																												
Amhara	515.68	503.52	349.57	196.48	294.02																																										
Anuak	356.86	330.81	863.72	552.48	411.33	516.69																																										
Ari Potter	26.38	0.00	826.95	642.89	231.89	503.52	330.81																																									
Ari Smith	26.38	0.00	826.95	642.89	231.89	503.52	330.81	0.00																																								
Bena	7.47	28.14	841.21	666.25	244.04	522.57	356.56	28.14	28.14																																								
Basketo	82.85	64.77	766.81	578.38	177.40	439.70	308.12	64.77	64.77	88.33																																							
Bume	5.67	31.97	836.44	664.72	239.03	519.32	362.21	31.97	31.97	7.98	87.86																																						
Bench	195.94	169.66	817.40	563.19	282.62	470.42	162.08	169.66	169.66	196.25	148.23	201.41																																					
Burji	116.54	134.08	769.45	647.86	185.76	479.73	447.75	134.08	134.08	122.12	139.83	114.15	286.79																																				
Busa	71.92	90.56	793.71	651.33	199.90	491.69	412.14	90.56	90.56	77.49	108.07	69.52	250.18	44.64																																				
Dizi	135.31	113.19	875.92	645.61	303.79	535.55	235.70	113.19	113.19	132.45	131.59	139.66	93.62	246.75	203.71																																				
Dime	6.93	27.16	840.48	665.30	243.35	521.71	355.76	27.16	27.16	1.01	87.35	8.16	195.40	122.21	77.56	131.90																																				
Dirasha	85.38	100.34	774.27	635.23	181.16	473.67	413.15	100.34	100.34	91.71	105.77	83.92	251.89	35.23	19.56	212.28	91.62																																				
Dasenech	2.67	29.02	834.21	661.67	236.88	516.61	359.53	29.02	29.02	8.05	84.65	3.28	198.60	114.65	70.01	137.81	7.82	83.83																																				
Dawuro	151.39	139.38	689.21	511.93	104.00	364.56	328.94	139.38	139.38	158.15	77.61	155.28	185.13	153.76	143.14	199.85	157.27	130.25	152.46																																				
Dorze	113.74	118.13	727.35	585.23	131.70	423.32	396.08	118.13	118.13	121.13	92.57	114.65	239.21	67.23	68.45	219.02	120.67	50.49	113.23	86.86																																				
Gobeze	66.87	86.91	801.07	657.09	206.90	498.31	411.17	86.91	86.91	72.04	109.08	64.07	249.11	50.40	7.54	200.08	72.19	27.06	64.81	147.91	75.29																																				
Gedeo	189.86	197.76	675.52	574.43	110.02	396.72	463.66	197.76	197.76	196.92	170.65	189.67	312.49	94.86	127.78	299.53	196.63	108.93	188.87	136.09	80.80	135.25																																				
Gofa	89.18	97.96	756.01	611.34	160.43	451.45	396.76	97.96	97.96	96.33	88.71	89.28	236.76	52.50	40.36	205.51	95.99	24.49	88.29	105.84	28.78	46.87	100.79																																				
Ganjule	83.41	96.23	769.32	627.12	174.89	466.66	405.20	96.23	96.23	90.11	97.43	82.56	244.21	42.59	25.07	207.04	89.93	9.26	82.11	121.20	43.39	32.03	107.80	15.81																																				
Gamo	99.53	104.77	741.20	595.31	144.92	435.41	391.15	104.77	104.77	106.90	84.72	100.36	232.67	63.35	56.48	207.96	106.45	40.07	98.98	91.41	14.33	62.91	93.20	16.13	31.82																																				
Genta	92.01	102.12	756.63	614.92	161.99	453.85	402.93	102.12	102.12	99.02	94.86	91.76	242.85	46.34	37.95	210.63	98.74	20.31	90.97	110.88	30.54	44.92	97.91	6.22	12.90	19.88																																				
Gurage	302.80	291.21	541.77	372.64	95.55	212.91	382.75	291.21	291.21	309.71	228.12	306.41	286.67	275.31	281.64	334.88	308.86	264.51	303.71	151.91	214.07	287.92	205.54	241.30	256.88	225.17	244.31																																				
Gewada	66.46	90.71	823.50	679.99	230.06	521.64	420.39	90.71	90.71	69.54	124.88	62.09	258.51	63.58	30.17	201.73	70.00	49.29	63.93	169.89	98.56	23.33	154.69	70.20	55.20	86.24	68.09	311.18																																				
Hadiya	246.55	237.62	589.65	434.91	39.91	271.76	381.57	237.62	237.62	253.71	177.37	249.73	262.72	213.01	220.39	295.59	252.91	202.87	247.22	99.80	152.37	226.85	146.41	180.05	195.47	163.95	182.78	62.30	250.17																																				
Hamer	1.64	27.87	833.60	660.67	236.31	515.79	358.43	27.87	27.87	8.16	83.56	4.45	197.47	115.09	70.46	136.95	7.79	84.08	1.20	151.58	112.97	65.36	188.84	88.21	82.23	98.74	90.96	302.89	64.82	246.48																																				
Kefa	192.68	167.01	761.07	516.17	230.81	415.04	190.21	167.01	167.01	195.15	124.95	198.35	57.21	263.46	232.93	129.67	194.18	230.27	195.18	139.66	207.30	233.57	273.48	211.22	221.64	203.70	217.44	229.57	247.50	208.06	193.98																																			
Kembata	210.52	203.74	623.57	474.56	32.28	310.91	382.63	203.74	203.74	217.82	146.92	213.35	250.84	173.30	180.85	271.89	217.07	163.13	211.00	72.04	112.65	187.41	112.29	140.56	155.87	124.51	143.12	102.09	210.75	39.89	210.33	199.81																																		
Konta	164.12	144.73	704.63	498.55	144.53	368.30	276.09	144.73	144.73	169.51	81.27	169.11	138.09	197.11	177.39	171.92	168.52	168.60	165.91	53.24	133.07	180.51	189.27	145.60	159.34	133.84	151.44	163.07	199.88	127.04	164.82	88.10	112.81																																	
Kore	98.10	113.15	767.28	634.31	176.70	470.20	423.67	113.15	113.15	104.34	115.73	96.51	262.80	24.10	29.48	224.88	104.28	12.85	96.51	133.53	49.35	36.86	98.46	29.16	18.64	42.12	23.25	262.58	56.88	200.54	96.78	239.49	160.66	174.69																																	
Konso	78.62	99.10	799.04	660.88	207.03	499.71	422.95	99.10	99.10	83.51	119.61	75.56	260.92	40.33	11.65	212.24	83.71	26.05	76.48	153.89	76.52	12.21	129.58	49.54	33.75	65.57	46.15	290.46	25.14	228.89	77.07	244.40	189.17	188.91	31.90																																	
Manja	159.72	145.88	685.61	501.66	107.47	357.69	317.79	145.88	145.88	166.24	82.21	163.89	177.25	167.55	155.71	198.33	165.33	143.39	160.96	14.02	100.77	160.22	148.73	119.08	134.27	104.95	124.26	145.91	181.84	97.89	160.03	129.51	75.19	41.72	147.05	166.62																																	
Mejenger	214.75	188.39	808.43	547.44	285.69	460.31	145.88	188.39	188.39	215.39	162.33	220.29	21.03	301.87	266.60	114.37	214.52	267.27	217.40	191.79	251.57	265.91	322.53	251.03	259.31	245.92	257.18	282.68	276.26	263.20	216.25	55.17	254.46	142.00	277.80	277.55	182.68																																	
Mashile	68.22	90.31	809.92	667.33	216.47	508.24	417.30	90.31	90.31	72.56	117.39	64.69	255.22	53.11	16.59	202.98	72.85	35.67	65.93	158.10	85.04	10.26	141.99	56.83	41.65	72.93	54.54	298.03	13.63	236.88	66.63	241.37	197.38	190.08	43.74	12.90	170.35	272.40																																	
Maale	6.73	21.09	835.89	659.24	239.01	516.27	350.75	21.09	21.09	7.39	81.15	11.50	190.07	122.90	78.33	128.60	6.38	91.32	9.22	151.75	117.87	73.42	194.85	94.06	89.00	103.77	97.16	303.47	73.16	247.89	8.36	188.05	212.37	162.26	104.09	85.23	159.58	209.04	74.93																																	
Nuer	436.38	411.41	955.02	628.09	513.09	611.25	102.86	411.41	411.41	434.84	398.54	441.35	250.35	536.23	497.73	305.06	434.16	501.11	439.02	427.40	489.29	495.82	560.31	486.92	493.77	483.01	492.97	485.42	502.12	484.06	438.01	287.75	483.86	375.26	512.41	507.92	416.94	238.05	500.83	429.89																																	
Oromo	244.66	229.19	614.69	416.23	104.15	277.57	317.18	229.19	229.19	250.99	164.48	248.99	211.50	243.12	238.22	261.23	250.05	223.95	246.00	96.04	176.42	243.35	197.06	199.48	215.21	184.05	203.92	75.41	265.65	67.29	245.04	154.54	86.07	90.89	225.43	248.57	85.43	208.29	253.58	244.16	419.98																																	
Shekecho	205.46	179.12	791.62	536.02	266.28	444.11	160.83	179.12	179.12	206.71	147.80	211.08	27.27	287.62	253.78	116.98	205.80	253.38	208.08	173.13	235.35	253.51	304.92	236.20	245.19	230.33	242.40	263.84	265.08	243.72	206.91	35.67	235.11	122.81	263.52	264.94	163.72	19.50	260.46	200.07	255.39	189.22																																	
Sidama	216.08	219.82	636.42	535.15	74.51	355.93	458.08	219.82	219.82	223.43	181.78	216.76	314.09	133.03	160.83	313.29	222.99	141.34	215.48	130.02	102.44	168.36	41.06	128.00	138.28	116.56	126.68	169.28	189.45	113.51	215.27	269.60	85.58	182.10	132.58	164.43	140.38	321.67	176.20	220.30	557.25	171.48	303.11																																	
Sheko	202.47	176.15	789.37	535.29	262.57	442.07	164.64	176.15	176.15	203.82	144.13	208.10	28.59	283.95	250.28	116.26	202.90	249.76	205.08	169.23	231.51	250.07	301.01	232.48	241.54	226.53	238.68	260.82	261.79	240.21	203.91	32.15	231.36	119.00	259.85	261.46	159.87	23.13	257.07	197.13	259.31	186.10	3.92	299.22																																	
Somali	742.71	744.97	287.27	550.26	521.77	415.29	891.36	744.97	744.97	750.07	697.55	743.41	795.35	648.17	684.27	825.56	749.62	665.36	742.13	625.73	628.98	691.73	556.50	654.44	663.62	643.17	652.69	512.60	710.54	533.80	741.91	739.23	553.72	660.81	654.91	685.59	628.13	793.64	698.21	746.81	992.78	585.53	774.42	526.65	771.16																																	
Tigrayan	393.26	379.40	472.88	277.51	188.52	127.58	406.23	379.40	379.40	399.90	314.96	397.23	344.59	370.63	376.37	407.97	399.00	359.52	394.38	241.97	309.17	382.49	298.01	336.08	351.75	319.96	339.28	95.35	405.67	157.63	393.49	288.56	197.44	241.37	357.85	385.41	233.78	335.62	392.66	393.32	505.35	151.15	318.74	259.75	316.50	490.96																																	
Wolayta	109.90	109.51	725.28	572.53	127.76	414.95	375.15	109.51	109.51	117.37	74.39	111.84	219.17	86.44	78.81	203.87	116.75	63.67	109.90	67.58	21.47	84.66	96.32	39.23	55.01	23.95	43.72	203.58	107.77	143.19	109.42	186.05	104.49	111.90	66.00	88.62	81.22	231.02	94.87	112.98	468.93	160.27	214.54	111.33	210.68	635.47	297.91																																	
Yem	255.84	242.38	594.28	409.03	86.55	261.41	341.91	242.38	242.38	262.49	178.31	259.82	235.81	242.70	242.44	282.13	261.60	226.82	256.97	104.60	177.53	248.10	187.49	202.67	218.48	186.76	206.53	52.76	270.96	47.19	256.08	179.06	75.26	110.38	226.88	252.20	96.53	233.12	258.35	255.95	444.74	25.19	213.97	158.08	210.78	560.57	137.41	163.65																																	
Zayse	89.69	100.40	759.74	617.82	165.12	456.94	403.25	100.40	100.40	96.64	95.14	89.31	242.92	45.30	34.83	209.51	96.38	17.48	88.59	113.22	33.63	41.78	100.38	7.56	9.77	22.60	3.13	247.31	64.97	185.82	88.61	218.21	146.18	153.14	21.69	43.17	126.51	257.45	51.42	94.93	492.94	206.54	242.82	129.56	239.12	655.44	342.25	46.28	209.34

Supplementary figure S6 Population pairwise distances (Fst) based on counts of donor's first language. Significant distances are highlighted in pink ($p < 0.05$). Basketo are also known as Masketo and Bume are also known as Nyangagtom.

	Ari	Ari Cultivator	Afar	Agew, Eastern	Alaba	Amhara	Anuak	Ari Potter	Ari Smith	Bena	Basketo	Bume	Bench	Burji	Busa	Dizi	Dime	Dirasha	Dasenech	Dawuro	Dorze	Gobeze	Gedeo	Gofa	Ganjule	Gamo	Genta	Gurage	Gewada	Hadiya	Hamer	Kefa	Kembata	Konta	Kore	Konso	Manja	Mejenger	Mashile	Maale	Nuer	Oromo	Shekecho	Sidama	Sheko	Somali	Tigrayan	Wolayta	Yem	Zayse
Ari	0.00																																																	
Ari Cultivator	0.05	0.00																																																
Afar	0.88	0.94	0.00																																															
Agew, Eastern	0.58	0.69	0.68	0.00																																													
Alaba	0.82	0.90	0.93	0.63	0.00																																												
Amhara	0.92	0.96	0.98	0.60	0.95	0.00																																											
Anuak	0.88	0.94	0.99	0.68	0.93	0.98	0.00																																										
Ari Potter	0.13	0.03	0.99	0.69	0.94	0.99	1.00	0.00																																									
Ari Smith	0.09	0.02	0.99	0.64	0.91	0.98	1.00	0.00	0.00																																								
Bena	0.76	0.90	0.95	0.54	0.85	0.97	0.97	0.97	0.94	0.00																																							
Basketo	0.85	0.92	0.96	0.65	0.90	0.97	0.96	0.97	0.95	0.91	0.00																																						
Bume	0.80	0.92	0.97	0.60	0.88	0.98	0.98	0.99	0.97	0.85	0.93	0.00																																					
Bench	0.88	0.94	0.98	0.68	0.92	0.98	0.98	0.98	0.98	0.94	0.95	0.96	0.00																																				
Burji	0.68	0.79	0.78	0.53	0.72	0.88	0.78	0.80	0.73	0.62	0.75	0.67	0.78	0.00																																				
Busa	0.74	0.84	0.84	0.59	0.79	0.91	0.85	0.86	0.81	0.72	0.82	0.76	0.84	0.65	0.00																																				
Dizi	0.78	0.87	0.88	0.59	0.82	0.92	0.88	0.89	0.85	0.78	0.85	0.81	0.88	0.69	0.75	0.00																																				
Dime	0.80	0.91	0.95	0.59	0.87	0.97	0.96	0.97	0.94	0.83	0.91	0.88	0.94	0.67	0.76	0.81	0.00																																				
Dirasha	0.62	0.76	0.74	0.43	0.68	0.80	0.75	0.76	0.68	0.55	0.71	0.61	0.74	0.53	0.60	0.64	0.61	0.00																																				
Dasenech	0.78	0.90	0.94	0.56	0.85	0.96	0.96	0.96	0.93	0.78	0.90	0.85	0.94	0.64	0.73	0.78	0.83	0.57	0.00																																				
Dawuro	0.90	0.94	0.96	0.74	0.93	0.97	0.96	0.97	0.96	0.94	0.95	0.95	0.96	0.84	0.88	0.90	0.94	0.81	0.93	0.00																																		
Dorze	0.84	0.91	0.95	0.64	0.89	0.96	0.95	0.96	0.94	0.88	0.92	0.91	0.94	0.73	0.80	0.84	0.89	0.69	0.88	0.94	0.00																																		
Gobeze	0.71	0.82	0.82	0.56	0.76	0.90	0.82	0.84	0.78	0.68	0.79	0.72	0.82	0.61	0.68	0.72	0.72	0.57	0.70	0.86	0.77	0.00																																	
Gedeo	0.88	0.94	0.98	0.68	0.93	0.98	0.99	0.99	0.99	0.96	0.96	0.97	0.98	0.78	0.85	0.88	0.95	0.75	0.95	0.96	0.95	0.82	0.00																																
Gofa	0.57	0.73	0.71	0.30	0.64	0.52	0.71	0.73	0.65	0.49	0.68	0.57	0.71	0.51	0.58	0.59	0.57	0.37	0.52	0.79	0.66	0.55	0.72	0.00																															
Ganjule	0.88	0.94	0.99	0.68	0.93	0.98	1.00	1.00	1.00	0.97	0.97	0.99	0.98	0.78	0.85	0.88	0.96	0.75	0.96	0.96	0.95	0.82	0.99	0.71	0.00																														
Gamo	0.61	0.73	0.72	0.33	0.67	0.40	0.72	0.74	0.68	0.56	0.70	0.63	0.72	0.56	0.61	0.62	0.62	0.42	0.58	0.78	0.68	0.59	0.73	0.15	0.72	0.00																													
Genta	0.65	0.78	0.77	0.48	0.71	0.84	0.77	0.79	0.71	0.59	0.74	0.65	0.77	0.56	0.63	0.67	0.65	0.49	0.61	0.83	0.72	0.60	0.77	0.43	0.77	0.47	0.00																												
Gurage	0.50	0.65	0.62	0.25	0.56	0.41	0.63	0.64	0.57	0.42	0.59	0.50	0.63	0.45	0.51	0.52	0.49	0.31	0.45	0.71	0.58	0.48	0.63	0.12	0.63	0.14	0.37	0.00																											
Gewada	0.88	0.94	0.98	0.68	0.93	0.98	0.99	0.99	0.99	0.95	0.96	0.97	0.98	0.78	0.84	0.88	0.95	0.74	0.94	0.96	0.95	0.82	0.98	0.71	0.99	0.72	0.77	0.62	0.00																									
Hadiya	0.73	0.83	0.83	0.55	0.78	0.89	0.84	0.85	0.80	0.70	0.81	0.75	0.83	0.64	0.71	0.74	0.75	0.58	0.72	0.87	0.79	0.68	0.84	0.53	0.84	0.57	0.62	0.46	0.83	0.00																								
Hamer	0.73	0.88	0.92	0.49	0.82	0.94	0.93	0.94	0.89	0.68	0.87	0.78	0.91	0.60	0.70	0.75	0.77	0.51	0.72	0.92	0.85	0.65	0.92	0.43	0.93	0.50	0.56	0.37	0.92	0.67	0.00																							
Kefa	0.79	0.88	0.89	0.60	0.83	0.93	0.90	0.90	0.87	0.79	0.86	0.83	0.89	0.69	0.76	0.79	0.82	0.64	0.80	0.91	0.85	0.73	0.89	0.60	0.90	0.63	0.67	0.53	0.89	0.75	0.76	0.00																						
Kembata	0.79	0.88	0.90	0.60	0.84	0.93	0.90	0.91	0.87	0.80	0.87	0.83	0.89	0.69	0.76	0.79	0.82	0.65	0.80	0.91	0.86	0.73	0.90	0.60	0.90	0.63	0.67	0.53	0.90	0.75	0.77	0.80	0.00																						
Konta	0.88	0.94	0.99	0.68	0.93	0.98	1.00	1.00	1.00	0.97	0.96	0.98	0.98	0.78	0.85	0.88	0.96	0.74	0.96	0.96	0.95	0.82	0.99	0.71	1.00	0.72	0.77	0.62	0.99	0.84	0.93	0.89	0.90	0.00																						
Kore	0.64	0.77	0.75	0.47	0.69	0.83	0.75	0.77	0.69	0.57	0.72	0.63	0.75	0.52	0.61	0.65	0.63	0.47	0.59	0.82	0.70	0.58	0.75	0.43	0.75	0.48	0.52	0.37	0.75	0.60	0.54	0.65	0.66	0.75	0.00																					
Konso	0.68	0.80	0.80	0.50	0.74	0.85	0.80	0.81	0.75	0.64	0.77	0.69	0.80	0.59	0.66	0.70	0.69	0.53	0.66	0.85	0.75	0.64	0.80	0.47	0.80	0.51	0.57	0.40	0.80	0.65	0.61	0.70	0.71	0.80	0.55	0.00																				
Manja	0.90	0.95	0.99	0.71	0.94	0.98	0.99	0.99	0.99	0.97	0.97	0.98	0.98	0.81	0.87	0.90	0.96	0.78	0.96	0.96	0.96	0.85	0.99	0.76	0.99	0.75	0.80	0.67	0.99	0.86	0.94	0.91	0.92	0.99	0.79	0.83	0.00																			
Mejenger	0.89	0.95	0.99	0.68	0.93	0.98	1.00	1.00	1.00	0.97	0.97	0.99	0.98	0.79	0.85	0.89	0.96	0.75	0.96	0.96	0.96	0.83	0.99	0.72	1.00	0.73	0.78	0.63	0.99	0.84	0.94	0.90	0.90	1.00	0.76	0.80	0.99	0.00																		
Mashile	0.77	0.86	0.87	0.61	0.82	0.92	0.87	0.88	0.84	0.77	0.84	0.80	0.87	0.67	0.74	0.78	0.80	0.64	0.78	0.89	0.83	0.69	0.87	0.61	0.87	0.64	0.67	0.54	0.87	0.74	0.74	0.78	0.79	0.87	0.64	0.70	0.89	0.88	0.00																	
Maale	0.81	0.90	0.92	0.63	0.86	0.95	0.93	0.93	0.90	0.84	0.89	0.87	0.92	0.72	0.79	0.82	0.86	0.68	0.84	0.92	0.88	0.76	0.92	0.64	0.93	0.67	0.70	0.56	0.92	0.78	0.81	0.83	0.83	0.92	0.69	0.74	0.93	0.93	0.81	0.00															
Nuer	0.89	0.95	0.99	0.68	0.94	0.99	1.00	1.00	1.00	0.97	0.97	0.99	0.98	0.79	0.85	0.89	0.96	0.75	0.96	0.96	0.96	0.83	0.99	0.72	1.00	0.73	0.78	0.63	0.99	0.84	0.94	0.90	0.91	1.00	0.76	0.81	0.99	1.00	0.88	0.93	0.00													
Oromo	0.67	0.78	0.77	0.48	0.72	0.81	0.77	0.79	0.73	0.62	0.75	0.67	0.77	0.49	0.65	0.68	0.67	0.48	0.64	0.82	0.73	0.58	0.77	0.43	0.77	0.48	0.55	0.37	0.77	0.63	0.59	0.68	0.68	0.77	0.47	0.58	0.80	0.78	0.65	0.72	0.78	0.00											
Shekecho	0.83	0.91	0.94	0.64	0.88	0.96	0.94	0.95	0.93	0.87	0.91	0.90	0.93	0.74	0.80	0.84	0.88	0.70	0.87	0.93	0.90	0.78	0.94	0.66	0.94	0.68	0.72	0.58	0.94	0.79	0.84	0.85	0.85	0.94	0.70	0.75	0.95	0.94	0.83	0.88	0.94	0.73	0.00									
Sidama	0.77	0.86	0.88	0.59	0.82	0.92	0.88	0.89	0.85	0.77	0.85	0.81	0.87	0.68	0.75	0.78	0.80	0.63	0.78	0.90	0.84	0.72	0.88	0.58	0.88	0.61	0.66	0.51	0.88	0.73	0.74	0.79	0.79	0.88	0.64	0.69	0.90	0.88	0.77	0.82	0.89	0.67	0.83	0.00							
Sheko	0.88	0.94	0.98	0.67	0.93	0.98	0.99	0.99	0.99	0.95	0.96	0.97	0.98	0.78	0.84	0.88	0.95	0.74	0.94	0.96	0.95	0.82	0.98	0.71	0.99	0.72	0.76	0.62	0.98	0.83	0.92	0.89	0.89	0.99	0.75	0.79	0.99	0.99	0.87	0.92	0.99	0.77	0.93	0.87	0.00						
Somali	0.81	0.89	0.92	0.62	0.86	0.95	0.92	0.93	0.90	0.83	0.89	0.86	0.91	0.71	0.78	0.81	0.85	0.66	0.83	0.92	0.87	0.75	0.92	0.62	0.92	0.65	0.69	0.54	0.92	0.76	0.79	0.82	0.83	0.92	0.67	0.72	0.93	0.92	0.80	0.85	0.93	0.70	0.87	0.81	0.91	0.00					
Tigrayan	0.62	0.79	0.78	0.34	0.70	0.72	0.78	0.80	0.71	0.53	0.74	0.62	0.78	0.54	0.62	0.64	0.62	0.41	0.57	0.84	0.72	0.58	0.79	0.24	0.79	0.28	0.47	0.18	0.78	0.58	0.48	0.66	0.66	0.78	0.46	0.51	0.82	0.79	0.66	0.70	0.79	0.48	0.72	0.63	0.78	0.68	0.00				
Wolayta	0.59	0.74	0.71	0.39	0.65	0.75	0.72	0.74	0.65	0.51	0.69	0.58	0.72	0.51	0.58	0.61	0.58	0.42	0.54	0.79	0.67	0.55	0.72	0.31	0.72	0.34	0.46	0.26	0.71	0.55	0.47	0.62	0.62	0.71	0.45	0.49	0.76	0.72	0.62	0.65	0.73	0.48	0.67	0.60	0.71	0.63	0.35	0.00			
Yem	0.87	0.94	0.98	0.67	0.92	0.98	0.99	0.99	0.99	0.95	0.96	0.97	0.97	0.77	0.84	0.88	0.95	0.74	0.94	0.96	0.94	0.81	0.98	0.70	0.99	0.72	0.76	0.62	0.98	0.83	0.92	0.89	0.89	0.99	0.74	0.79	0.99	0.99	0.86	0.92	0.99	0.76	0.93	0.87	0.98	0.91	0.77	0.71	0.00	
Zayse	0.79	0.88	0.89	0.60	0.84	0.93	0.90	0.91	0.87	0.79	0.87	0.83	0.89	0.69	0.75	0.79	0.82	0.64	0.80	0.91	0.85	0.73	0.90	0.60	0.90	0.63	0.67	0.52	0.89	0.74	0.76	0.80	0.81	0.90	0.65	0.70	0.91	0.90	0.79	0.83	0.90	0.68	0.85	0.79	0.89	0.82	0.65	0.61	0.89	0.00

Supplementary figure S7 Population pairwise distances (Fst) based on counts of donor's first language (linguistic group). Significant distances are highlighted in pink ($p < 0.05$). Basketo are also known as Masketo and Bume are also known as Nyangagtom.

	Ari	Ari Cultivator	Afar	Agew, Eastern	Alaba	Amhara	Anuak	Ari Potter	Ari Smith	Bena	Basketo	Bume	Bench	Burji	Busa	Dizi	Dime	Dirasha	Dasenech	Dawuro	Dorze	Gobeze	Gedeo	Gofa	Ganjule	Gamo	Genta	Gurage	Gewada	Hadiya	Hamer	Kefa	Kembata	Konta	Kore	Konso	Manja	Mejenger	Mashile	Maale	Nuer	Oromo	Shekecho	Sidama	Sheko	Somali	Tigrayan	Wolayta	Yem	Zayse
Ari	-1.00																																																
Ari Cultivator	0.05	-1.00																																															
Afar	0.88	0.94	-1.00																																													
Agew, Eastern	0.58	0.69	0.34	-1.00																																											
Alaba	0.82	0.90	0.04	0.26	-1.00																																										
Amhara	0.92	0.96	0.98	0.59	0.95	-1.00																																									
Anuak	0.88	0.94	0.99	0.68	0.93	0.98	-1.00																																								
Ari Potter	0.13	0.03	0.99	0.69	0.94	0.99	1.00	-1.00																																							
Ari Smith	0.09	0.02	0.99	0.64	0.91	0.98	1.00	0.00	-1.01																																						
Bena	-0.04	0.05	0.95	0.54	0.85	0.97	0.97	0.41	0.23	-1.03																																					
Basketo	0.06	0.00	0.97	0.66	0.91	0.97	0.97	0.02	0.01	0.11	-1.00																																				
Bume	0.80	0.92	0.97	0.60	0.88	0.98	0.09	0.99	0.97	0.85	0.94	-1.02																																			
Bench	0.11	0.01	0.98	0.68	0.93	0.98	0.99	0.00	-0.01	0.27	0.00	0.97	-1.00																																		
Burji	0.88	0.94	0.00	0.35	0.05	0.98	0.99	0.99	0.99	0.96	0.97	0.97	0.98	-1.00																																	
Busa	0.78	0.87	0.07	0.29	0.04	0.94	0.91	0.91	0.87	0.80	0.87	0.85	0.90	0.06	-1.00																																
Dizi	-0.01	0.04	0.88	0.59	0.82	0.92	0.88	0.11	0.08	-0.04	0.05	0.81	0.09	0.88	0.79	-1.00																															
Dime	0.00	0.00	0.95	0.59	0.87	0.97	0.96	0.13	0.06	-0.02	0.01	0.88	0.07	0.95	0.82	0.00	-1.01																														
Dirasha	0.67	0.80	0.24	0.06	0.12	0.84	0.80	0.81	0.74	0.62	0.77	0.68	0.80	0.25	0.15	0.69	0.68	-1.00																													
Dasenech	0.78	0.90	0.15	0.17	-0.01	0.96	0.96	0.96	0.93	0.78	0.92	0.85	0.95	0.17	0.04	0.78	0.83	0.04	-1.02																												
Dawuro	0.12	0.01	0.97	0.75	0.94	0.97	0.97	0.01	0.00	0.20	0.00	0.96	0.00	0.97	0.91	0.10	0.04	0.86	0.95	-1.00																											
Dorze	0.05	-0.01	0.95	0.65	0.90	0.97	0.96	0.03	0.01	0.06	-0.01	0.92	0.01	0.96	0.86	0.04	0.00	0.75	0.89	0.00	-1.00																										
Gobeze	0.89	0.95	0.00	0.36	0.06	0.98	1.00	1.00	1.00	0.97	0.97	0.99	0.99	0.00	0.08	0.89	0.96	0.26	0.24	0.97	0.96	-1.00																									
Gedeo	0.89	0.95	0.00	0.36	0.06	0.99	1.00	1.00	1.00	0.97	0.98	0.99	0.99	0.00	0.09	0.89	0.96	0.27	0.24	0.97	0.97	0.00	-1.00																									
Gofa	0.35	0.57	0.75	0.34	0.68	0.59	0.75	0.59	0.49	0.28	0.52	0.62	0.57	0.75	0.66	0.37	0.38	0.47	0.57	0.66	0.50	0.76	0.76	-1.00																								
Ganjule	0.12	0.03	0.99	0.68	0.93	0.98	1.00	0.00	0.00	0.37	0.02	0.99	0.00	0.99	0.90	0.10	0.11	0.80	0.96	0.01	0.02	1.00	1.00	0.57	-1.00																							
Gamo	0.43	0.60	0.72	0.33	0.67	0.40	0.73	0.62	0.55	0.38	0.57	0.63	0.60	0.73	0.66	0.45	0.46	0.48	0.59	0.68	0.55	0.73	0.73	0.01	0.60	-1.00																						
Genta	0.00	0.11	0.83	0.53	0.77	0.89	0.84	0.18	0.13	-0.03	0.11	0.74	0.16	0.83	0.73	0.01	0.03	0.61	0.70	0.20	0.09	0.84	0.84	0.27	0.17	0.36	-1.00																				
Gurage	0.87	0.94	0.98	0.48	0.92	0.00	0.98	0.99	0.98	0.94	0.96	0.96	0.98	0.98	0.91	0.87	0.94	0.74	0.94	0.97	0.95	0.99	0.99	0.44	0.98	0.29	0.81	-1.00																		
Gewada	0.88	0.94	-0.01	0.34	0.04	0.98	0.99	0.99	0.99	0.95	0.97	0.97	0.98	0.00	0.07	0.88	0.95	0.24	0.15	0.97	0.95	0.00	0.00	0.75	0.99	0.72	0.83	0.98	-1.00																
Hadiya	0.74	0.84	0.13	0.15	0.03	0.90	0.85	0.86	0.82	0.73	0.83	0.77	0.85	0.14	0.07	0.75	0.76	0.02	-0.01	0.89	0.81	0.15	0.15	0.58	0.85	0.58	0.69	0.83	0.13	-1.00														
Hamer	0.01	0.21	0.92	0.50	0.82	0.94	0.93	0.47	0.29	-0.03	0.25	0.78	0.39	0.92	0.77	0.01	0.06	0.59	0.72	0.40	0.20	0.94	0.94	0.20	0.43	0.30	-0.02	0.90	0.92	0.70	-1.02												
Kefa	0.00	0.02	0.89	0.60	0.83	0.93	0.90	0.08	0.05	-0.03	0.03	0.83	0.07	0.89	0.79	0.00	-0.01	0.69	0.79	0.07	0.02	0.90	0.90	0.40	0.08	0.47	0.02	0.89	0.89	0.76	0.04	-1.00											
Kembata	0.80	0.88	0.06	0.23	-0.01	0.94	0.91	0.92	0.88	0.81	0.89	0.84	0.91	0.07	0.03	0.80	0.83	0.09	-0.02	0.92	0.87	0.08	0.08	0.65	0.91	0.64	0.74	0.90	0.06	0.01	0.78	0.81	-1.00										
Konta	0.86	0.94	0.99	0.47	0.93	0.00	1.00	1.00	1.00	0.96	0.97	0.98	0.99	0.99	0.91	0.87	0.96	0.73	0.95	0.97	0.96	1.00	1.00	0.41	1.00	0.27	0.80	0.00	0.99	0.82	0.91	0.89	0.90	-1.00									
Kore	0.07	0.20	0.71	0.42	0.64	0.84	0.76	0.24	0.17	0.03	0.18	0.63	0.22	0.71	0.59	0.08	0.09	0.47	0.54	0.29	0.15	0.72	0.72	0.22	0.22	0.32	0.04	0.74	0.71	0.56	0.02	0.08	0.61	0.71	-1.00								
Konso	0.69	0.81	0.17	0.10	0.07	0.86	0.81	0.82	0.76	0.65	0.78	0.71	0.81	0.18	0.10	0.70	0.70	0.00	0.01	0.86	0.77	0.19	0.20	0.51	0.81	0.52	0.63	0.78	0.17	0.00	0.62	0.70	0.04	0.76	0.49	-1.00							
Manja	0.15	0.03	0.99	0.71	0.95	0.99	1.00	0.00	0.00	0.48	0.03	0.99	0.00	0.99	0.92	0.13	0.16	0.84	0.97	0.01	0.03	1.00	1.00	0.63	0.00	0.64	0.21	0.99	0.99	0.88	0.53	0.10	0.93	1.00	0.27	0.84	-1.00						
Mejenger	0.89	0.95	0.99	0.68	0.94	0.98	0.00	1.00	1.00	0.97	0.97	0.09	0.99	0.99	0.91	0.89	0.96	0.81	0.96	0.97	0.96	1.00	1.00	0.76	1.00	0.73	0.84	0.99	0.99	0.86	0.94	0.90	0.91	1.00	0.76	0.82	1.00	-1.00					
Mashile	0.90	0.95	0.00	0.37	0.07	0.99	1.00	1.00	1.00	0.97	0.98	0.99	0.99	0.00	0.09	0.90	0.97	0.28	0.26	0.97	0.97	0.00	0.00	0.77	1.00	0.74	0.85	0.99	0.00	0.16	0.94	0.91	0.09	1.00	0.74	0.21	1.00	1.00	-1.00				
Maale	0.04	-0.01	0.95	0.65	0.90	0.97	0.96	0.03	0.02	0.05	-0.01	0.92	0.02	0.95	0.86	0.03	-0.01	0.76	0.89	0.00	-0.01	0.96	0.96	0.50	0.03	0.55	0.09	0.95	0.95	0.82	0.19	0.02	0.87	0.96	0.16	0.77	0.04	0.96	0.95	-1.00			
Nuer	0.89	0.95	0.99	0.69	0.94	0.99	0.00	1.00	1.00	0.97	0.97	0.09	0.99	0.99	0.91	0.89	0.96	0.81	0.96	0.97	0.96	1.00	1.00	0.76	1.00	0.73	0.84	0.99	0.99	0.86	0.94	0.90	0.91	1.00	0.77	0.82	1.00	0.00	1.00	0.96	-1.00		
Oromo	0.67	0.78	0.23	0.05	0.12	0.81	0.77	0.79	0.73	0.62	0.75	0.67	0.78	0.24	0.16	0.68	0.67	-0.01	0.05	0.83	0.73	0.25	0.25	0.47	0.77	0.48	0.61	0.70	0.23	0.03	0.59	0.68	0.09	0.69	0.48	0.00	0.81	0.78	0.26	0.74	0.78	-1.00	
Shekecho	0.03	0.00	0.93	0.64	0.88	0.96	0.94	0.04	0.02	0.02	0.00	0.90	0.02	0.94	0.84	0.03	-0.01	0.74	0.87	0.01	-0.01	0.94	0.94	0.49	0.03	0.54	0.08	0.93	0.93	0.80	0.15	0.01	0.86	0.94	0.14	0.76	0.05	0.94	0.95	-0.01	0.94	0.73	-1.00
Sidama	0.77	0.86	0.09	0.19	0.01	0.92	0.88	0.89	0.85	0.77	0.86	0.81	0.88	0.10	0.05	0.78	0.80	0.05	-0.02	0.91	0.84	0.12	0.12	0.62	0.88	0.62	0.72	0.87	0.09	0.00	0.74	0.78	0.00	0.86	0.59	0.02	0.90	0.88	0.12	0.85	0.89	0.06	0.83	-1.00	
Sheko	0.10	0.01	0.98	0.68	0.93	0.98	0.99	0.00	-0.01	0.24	0.00	0.97	-0.01	0.98	0.89	0.09	0.06	0.79	0.94	0.00	0.01	0.99	0.99	0.56	0.00	0.59	0.15	0.98	0.98	0.85	0.36	0.06	0.90	0.99	0.21	0.80	0.00	0.99	0.99	0.01	0.99	0.77	0.02	0.87	-1.00	
Somali	0.81	0.90	0.04	0.25	-0.01	0.95	0.93	0.94	0.91	0.84	0.91	0.87	0.93	0.05	0.04	0.82	0.86	0.11	-0.01	0.94	0.89	0.07	0.07	0.67	0.93	0.66	0.76	0.92	0.04	0.03	0.81	0.83	-0.01	0.92	0.63	0.06	0.95	0.93	0.08	0.89	0.93	0.12	0.88	0.01	0.92	-1.00	
Tigrayan	0.83	0.93	0.98	0.42	0.90	-0.01	0.99	0.99	0.98	0.92	0.96	0.96	0.98	0.98	0.88	0.83	0.92	0.67	0.91	0.96	0.94	0.99	0.99	0.35	0.99	0.24	0.75	-0.01	0.98	0.78	0.85	0.86	0.87	0.01	0.66	0.71	0.99	0.99	0.99	0.94	0.99	0.64	0.92	0.83	0.98	0.90	-1.01	
Wolayta	0.10	0.29	0.76	0.44	0.70	0.79	0.77	0.34	0.26	0.05	0.27	0.64	0.32	0.77	0.67	0.12	0.14	0.52	0.60	0.40	0.24	0.77	0.78	0.10	0.32	0.19	0.05	0.68	0.76	0.62	0.01	0.15	0.67	0.65	0.05	0.55	0.38	0.77	0.79	0.24	0.78	0.53	0.23	0.65	0.31	0.69	0.59	-1.00	
Yem	0.10	0.01	0.98	0.67	0.92	0.98	0.99	0.00	-0.01	0.24	0.00	0.97	-0.01	0.98	0.89	0.08	0.06	0.79	0.94	0.00	0.01	0.99	0.99	0.55	0.00	0.59	0.15	0.98	0.98	0.84	0.35	0.06	0.90	0.99	0.20	0.80	0.00	0.99	0.99	0.01	0.99	0.76	0.02	0.87	-0.01	0.92	0.98	0.30	-1.00	
Zayse	0.00	0.02	0.90	0.61	0.84	0.94	0.91	0.10	0.06	-0.03	0.04	0.84	0.08	0.90	0.81	-0.01	-0.01	0.70	0.81	0.07	0.02	0.91	0.91	0.40	0.09	0.47	0.02	0.90	0.90	0.77	0.04	-0.01	0.82	0.90	0.09	0.72	0.12	0.91	0.92	0.02	0.91	0.69	0.01	0.80	0.07	0.84	0.87	0.14	0.07	-1.00

Supplementary figure S8 Population pairwise distances (Fst) based on counts of donor's religion. Significant distances are highlighted in pink ($p < 0.05$). Basketo are also known as Masketo and Bume are also known as Nyangagtom.

	Ari	Ari Cultivator	Afar	Agew, Eastern	Alaba	Amhara	Anuak	Ari Potter	Ari Smith	Bena	Basketo	Bume	Bench	Burji	Busa	Dizi	Dime	Dirasha	Dasenech	Dawuro	Dorze	Gobeze	Gedeo	Gofa	Ganjule	Gamo	Genta	Gurage	Gewada	Hadiya	Hamer	Kefa	Kembata	Konta	Kore	Konso	Manja	Mejenger	Mashile	Maale	Nuer	Oromo	Shekecho	Sidama	Sheko	Somali	Tigrayan	Wolayta	Yem	Zayse
Ari	-1.00																																																
Ari Cultivator	0.04	-1.00																																																
Afar	1.00	0.92	-1.00																																														
Agew, Eastern	0.02	0.02	0.96	-1.00																																												
Alaba	0.98	0.91	0.01	0.93	-1.00																																											
Amhara	0.10	0.08	0.77	0.08	0.76	-1.00																																											
Anuak	0.00	0.04	1.00	0.01	0.98	0.10	-1.00																																									
Ari Potter	0.35	0.28	0.76	0.39	0.74	0.24	0.34	-1.00																																								
Ari Smith	0.36	0.29	0.87	0.35	0.84	0.15	0.35	0.01	-1.01																																							
Bena	0.00	0.00	1.00	-0.02	0.97	0.06	0.00	0.22	0.15	-1.03																																						
Basketo	0.00	0.04	1.00	0.02	0.98	0.10	0.05	0.35	0.35	0.00	-1.00																																				
Bume	0.09	-0.02	0.99	0.00	0.98	0.05	0.09	0.19	0.12	-0.02	0.09	-1.02																																			
Bench	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.36	0.37	0.00	0.00	0.10	-1.00																																		
Burji	0.02	0.02	0.97	-0.01	0.95	0.06	0.02	0.31	0.27	-0.02	0.02	0.00	0.03	-1.00																																	
Busa	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.36	0.37	0.00	0.00	0.10	0.00	0.03	-1.00																																
Dizi	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.36	0.36	0.00	0.00	0.10	0.00	0.03	0.00	-1.00																															
Dime	0.00	0.02	1.00	0.00	0.97	0.08	0.00	0.28	0.22	0.00	0.00	0.02	0.00	0.01	0.00	0.00	-1.01																														
Dirasha	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.34	0.00	0.00	0.08	0.00	0.02	0.00	0.00	0.00	-1.00																													
Dasenech	0.00	0.01	1.00	0.00	0.97	0.08	0.00	0.24	0.19	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	-1.02																												
Dawuro	0.00	0.06	1.00	0.03	0.96	0.14	0.00	0.46	0.55	0.00	0.00	0.22	0.00	0.05	0.00	0.00	0.00	0.00	0.00	-1.00																									
Dorze	0.00	0.03	1.00	0.01	0.96	0.10	0.00	0.34	0.34	0.00	0.00	0.08	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	-1.00																								
Gobeze	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.35	0.36	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.00																							
Gedeo	0.16	0.11	0.82	0.13	0.80	0.00	0.16	0.20	0.13	0.07	0.16	0.06	0.17	0.09	0.17	0.17	0.11	0.15	0.09	0.27	0.15	0.16	-1.00																						
Gofa	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	-1.00																					
Ganjule	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	-1.00																			
Gamo	0.00	0.04	0.99	0.01	0.97	0.11	0.00	0.40	0.41	-0.03	0.00	0.03	0.00	0.01	0.00	0.00	-0.01	0.00	-0.02	0.01	0.00	0.00	0.17	0.00	0.00	-1.00																	
Genta	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	-1.00																
Gurage	0.03	0.02	0.95	0.00	0.93	0.05	0.03	0.31	0.26	-0.01	0.03	0.00	0.03	-0.01	0.03	0.04	0.01	0.03	0.01	0.06	0.03	0.03	0.08	0.03	0.03	0.02	0.03	-1.00															
Gewada	0.13	0.07	0.84	0.10	0.81	0.00	0.13	0.18	0.11	0.05	0.13	0.04	0.14	0.07	0.14	0.14	0.09	0.13	0.07	0.23	0.13	0.14	-0.01	0.13	0.13	0.15	0.13	0.06	-1.00													
Hadiya	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.36	0.37	0.00	0.00	0.10	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.04	0.14	-1.00													
Hamer	0.00	0.01	1.00	-0.01	0.97	0.07	0.00	0.24	0.18	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	-0.02	0.00	0.00	0.07	0.00	-1.02											
Kefa	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.35	0.36	0.00	0.00	0.09	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.14	0.00	0.00	-1.00										
Kembata	0.00	0.04	1.00	0.02	0.98	0.10	0.00	0.35	0.36	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	-1.00									
Konta	0.00	0.03	1.00	0.01	0.96	0.10	0.00	0.34	0.34	0.00	0.00	0.08	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	-1.00								
Kore	0.00	0.04	1.00	0.02	0.98	0.10	0.00	0.35	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.00	-1.00							
Konso	0.03	0.02	0.96	0.00	0.94	0.05	0.03	0.30	0.25	-0.01	0.03	0.00	0.03	-0.01	0.03	0.04	0.01	0.03	0.00	0.07	0.03	0.03	0.08	0.03	0.03	0.02	0.03	-0.01	0.06	0.04	0.00	0.03	0.03	0.03	0.03	-1.00						
Manja	0.85	0.85	0.87	0.86	0.85	0.72	0.85	0.43	0.58	0.79	0.85	0.77	0.86	0.82	0.85	0.86	0.81	0.85	0.80	0.91	0.84	0.85	0.72	0.85	0.85	0.87	0.85	0.82	0.71	0.98	0.79	0.85	0.85	0.84	0.85	0.81	-1.00					
Mejenger	0.00	0.04	1.00	0.02	0.98	0.10	0.00	0.35	0.36	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	-1.00				
Mashile	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.37	0.39	0.00	0.00	0.11	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.04	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.86	0.00	-1.00			
Maale	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.35	0.36	0.00	0.00	0.10	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	-1.00		
Nuer	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.35	0.36	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	0.00	-1.00	
Oromo	0.14	0.09	0.82	0.11	0.83	0.00	0.14	0.21	0.14	0.06	0.14	0.06	0.15	0.08	0.15	0.15	0.10	0.14	0.08	0.23	0.13	0.14	-0.01	0.14	0.14	0.15	0.14	0.07	-0.01	0.15	0.08	0.14	0.14	0.13	0.14	0.07	0.72	0.14	0.15	0.14	0.14	-1.00
Shekecho	0.00	0.04	1.00	0.02	0.98	0.11	0.00	0.36	0.37	0.00	0.00	0.10	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	0.00	0.00	0.14	-1.00
Sidama	0.10	0.05	0.86	0.05	0.86	0.00	0.10	0.24	0.16	0.03	0.10	0.03	0.10	0.04	0.10	0.11	0.05	0.10	0.05	0.17	0.09	0.10	0.01	0.10	0.10	0.10	0.10	0.02	0.00	0.10	0.04	0.10	0.10	0.09	0.10	0.03	0.75	0.10	0.11	0.10	0.10	0.00	0.10	-1.00				
Sheko	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	0.00	0.00	0.14	0.00	0.10	-1.00			
Somali	0.97	0.90	0.02	0.94	-0.01	0.75	0.97	0.72	0.82	0.95	0.97	0.94	0.97	0.94	0.97	0.97	0.96	0.97	0.96	0.99	0.97	0.97	0.78	0.97	0.97	0.97	0.97	0.92	0.80	0.97	0.95	0.97	0.97	0.97	0.97	0.93	0.84	0.97	0.98	0.97	0.97	0.78	0.97	0.85	0.97	-1.00			
Tigrayan	0.19	0.10	0.88	0.13	0.85	-0.01	0.18	0.20	0.13	0.07	0.19	0.06	0.20	0.09	0.20	0.20	0.11	0.18	0.09	0.34	0.18	0.19	-0.01	0.19	0.18	0.20	0.19	0.07	-0.01	0.20	0.08	0.19	0.19	0.18	0.19	0.07	0.73	0.19	0.21	0.19	0.19	-0.01	0.20	0.00	0.19	0.83	-1.01		
Wolayta	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	0.00	0.00	0.14	0.00	0.10	0.00	0.97	0.19	-1.00	
Yem	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	0.00	0.00	0.14	0.00	0.10	0.00	0.97	0.19	0.00	-1.00	
Zayse	0.00	0.04	1.00	0.01	0.98	0.10	0.00	0.34	0.35	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00	0.00	0.00	0.00	0.14	0.00	0.10	0.00	0.97	0.19	0.00	0.00	-1.00

References

Abecasis,G.R. and Cookson,W.O. (2000). GOLD--graphical overview of linkage disequilibrium. *Bioinformatics.* **16,** 182-183.

Abernethy,D.R. and Todd,E.L. (1985). Impairment of caffeine clearance by chronic use of low-dose oestrogen-containing oral contraceptives. *Eur. J. Clin. Pharmacol.* **28,** 425-428.

Agundez,J.A. (2004). Cytochrome P450 gene polymorphism and cancer. *Curr. Drug Metab.* **5,** 211-224.

Aklillu,E., Carrillo,J.A., Makonnen,E., Hellman,K., Pitarque,M., Bertilsson,L. and Ingelman-Sundberg,M. (2003). Genetic polymorphism of CYP1A2 in Ethiopians affecting induction and expression: characterization of novel haplotypes with single-nucleotide polymorphisms in intron 1. *Mol. Pharmacol.* **64,** 659-669.

Allorge,D., Chevalier,D., Lo-Guidice,J.M., Cauffiez,C., Suard,F., Baumann,P., Eap,C.B. and Broly,F. (2003). Identification of a novel splice-site mutation in the CYP1A2 gene. *Br. J. Clin. Pharmacol.* **56,** 341-344.

Andres,A.M., Clark,A.G., Shimmin,L., Boerwinkle,E., Sing,C.F. and Hixson,J.E. (2007). Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epidemiol.* **31,** 659-671.

Anthony,F., Combes,C., Astorga,C., Bertrand,B., Graziosi,G. and Lashermes,P. (2002). The origin of cultivated Coffea arabica L. varieties revealed by AFLP and SSR markers. *Theor. Appl. Genet.* **104,** 894-900.

Arnett,D.K., Claas,S.A. and Lynch,A.I. (2009). Has pharmacogenetics brought us closer to 'personalized medicine' for initial drug treatment of hypertension? *Curr. Opin. Cardiol.* **24,** 333-339.

Bailey,W.J., Fitch,D.H., Tagle,D.A., Czelusniak,J., Slightom,J.L. and Goodman,M. (1991). Molecular evolution of the psi eta-globin gene locus: gibbon phylogeny and the hominoid slowdown. *Mol. Biol. Evol.* **8,** 155-184.

Balogh,A., Klinger,G., Henschel,L., Borner,A., Vollanth,R. and Kuhnz,W. (1995). Influence of ethinylestradiol-containing combination oral contraceptives with gestodene or levonorgestrel on caffeine elimination. *Eur. J. Clin. Pharmacol.* **48,** 161-166.

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

Bates,D.W., Spell,N., Cullen,D.J., Burdick,E., Laird,N., Petersen,L.A., Small,S.D., Sweitzer,B.J. and Leape,L.L. (1997). The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *JAMA.* **277,** 307-311.

Behar,D.M., Thomas,M.G., Skorecki,K., Hammer,M.F., Bulygina,E., Rosengarten,D., Jones,A.L., Held,K., Moses,V., Goldstein,D., Bradman,N. and Weale,M.E. (2003). Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. *Am. J. Hum. Genet.* **73,** 768-779.

Bertilsson,L., Dahl,M.L., Sjoqvist,F., Aberg-Wistedt,A., Humble,M., Johansson,I., Lundqvist,E. and Ingelman-Sundberg,M. (1993). Molecular basis for rational megaprescribing in ultrarapid hydroxylators of debrisoquine. *Lancet.* **341,** 63.

Bertilsson,L., Carrillo,J.A., Dahl,M.L., Llerena,A., Alm,C., Bondesson,U., Lindstrom,L., Rodriguez,d.l.R., I, Ramos,S. and Benitez,J. (1994). Clozapine disposition covaries with CYP1A2 activity determined by a caffeine test. *Br. J. Clin. Pharmacol.* **38,** 471-473.

Betti,L., Balloux,F., Hanihara,T. and Manica,A. (2009). The relative role of drift and selection in shaping the human skull. *Am. J. Phys. Anthropol.*

Bonilla,C., Panguluri,R.K., Taliaferro-Smith,L., Argyropoulos,G., Chen,G., Adeyemo,A.A., Amoah,A., Owusu,S., Acheampong,J., Agyenim-Boateng,K., Eghan,B.A., Oli,J., Okafor,G., Abbiyesuku,F., Johnson,T., Rufus,T., Fasanmade,O., Chen,Y., Collins,F.S., Dunston,G.M., Rotimi,C. and Kittles,R.A. (2006). Agouti-related protein promoter variant associated with leanness and decreased risk for diabetes in West Africans. *Int. J. Obes. (Lond).* **30,** 715-721.

Boobis,A.R., Lynch,A.M., Murray,S., de la,T.R., Solans,A., Farre,M., Segura,J., Gooderham,N.J. and Davies,D.S. (1994). CYP1A2-catalyzed conversion of dietary heterocyclic amines to their proximate carcinogens is their major route of metabolism in humans. *Cancer Res.* **54,** 89-94.

Burnet,D.L., Elliott,L.D., Quinn,M.T., Plaut,A.J., Schwartz,M.A. and Chin,M.H. (2006). Preventing diabetes in the clinical setting. *J. Gen. Intern. Med.* **21,** 84-93.

Butler,M.A., Iwasaki,M., Guengerich,F.P. and Kadlubar,F.F. (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,** 7696-7700.

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

Carroll,S.B. (2003). Genetics and the making of Homo sapiens. *Nature.* **422,** 849-857.

Castorena-Torres,F., Mendoza-Cantu,A., de Leon,M.B., Cisneros,B., Zapata-Perez,O., Lopez-Carrillo,L., Salinas,J.E. and Albores,A. (2005). CYP1A2 phenotype and genotype in a population from the Carboniferous Region of Coahuila, Mexico. *Toxicol. Lett.* **156,** 331-339.

Cauchi,S., El Achhab,Y., Choquet,H., Dina,C., Krempler,F., Weitgasser,R., Nejjari,C., Patsch,W., Chikri,M., Meyre,D. and Froguel,P. (2007). TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *J. Mol. Med.* **85,** 777-782.

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

Chandak,G.R., Janipalli,C.S., Bhaskar,S., Kulkarni,S.R., Mohankrishna,P., Hattersley,A.T., Frayling,T.M. and Yajnik,C.S. (2007). Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population. *Diabetologia.* **50,** 63-67.

Chen,F.C. and Li,W.H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68,** 444-456.

Chen,G., Adeyemo,A., Zhou,J., Chen,Y., Huang,H., Doumatey,A., Lashley,K., Agyenim-Boateng,K., Eghan,B.A., Jr., Acheampong,J., Fasanmade,O., Johnson,T., Okafor,G., Oli,J., Amoah,A. and Rotimi,C. (2007). Genome-wide search for susceptibility genes to type 2 diabetes in West Africans: potential role of C-peptide. *Diabetes Res. Clin. Pract.* **78,** e1-e6.

Chen,G., Adeyemo,A.A., Zhou,J., Chen,Y., Doumatey,A., Lashley,K., Huang,H., Amoah,A., Agyenim-Boateng,K., Eghan,B.A., Jr., Okafor,G., Acheampong,J., Oli,J., Fasanmade,O., Johnson,T. and Rotimi,C. (2007). A genome-wide search for linkage to renal function phenotypes in West Africans with type 2 diabetes. *Am. J. Kidney Dis.* **49,** 394-400.

Chen,X., Wang,L., Zhi,L., Zhou,G., Wang,H., Zhang,X., Hao,B., Zhu,Y., Cheng,Z. and He,F. (2005). The G-113A polymorphism in CYP1A2 affects the caffeine metabolic ratio in a Chinese population. *Clin. Pharmacol. Ther.* **78,** 249-259.

Chevalier,D., Cauffiez,C., Allorge,D., Lo-Guidice,J.M., Lhermitte,M., Lafitte,J.J. and Broly,F. (2001). Five novel natural allelic variants-951A>C, 1042G>A (D348N), 1156A>T (I386F), 1217G>A (C406Y) and 1291C>T (C431Y)-of the human CYP1A2 gene in a French Caucasian population. *Hum. Mutat.* **17,** 355-356.

Chida,M., Yokoi,T., Fukui,T., Kinoshita,M., Yokota,J. and Kamataki,T. (1999). Detection of three genetic polymorphisms in the 5'-flanking region and intron 1 of human CYP1A2 in the Japanese population. *Jpn. J. Cancer Res.* **90,** 899-902.

Chung,I. and Bresnick,E. (1995). Regulation of the constitutive expression of the human CYP1A2 gene: cis elements and their interactions with proteins. *Mol. Pharmacol.* **47,** 677-685.

Chung,I. and Bresnick,E. (1997). Identification of positive and negative regulatory elements of the human cytochrome P4501A2 (CYP1A2) gene. *Arch. Biochem. Biophys.* **338,** 220-226.

Clark,A.G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7,** 111-122.

Cornelis,M.C., El Sohemy,A. and Campos,H. (2004). Genetic polymorphism of CYP1A2 increases the risk of myocardial infarction. *J. Med. Genet.* **41,** 758-762.

Coutts,R.T. and Urichuk,L.J. (1999). Polymorphic cytochromes P450 and drugs used in psychiatry. *Cell Mol. Neurobiol.* **19,** 325-354.

Damcott,C.M., Pollin,T.I., Reinhart,L.J., Ott,S.H., Shen,H., Silver,K.D., Mitchell,B.D. and Shuldiner,A.R. (2006). Polymorphisms in the transcription factor 7-like 2 (TCF7L2) gene are associated with type 2 diabetes in the Amish: replication and evidence for a role in both insulin secretion and insulin resistance. *Diabetes.* **55,** 2654-2659.

Dandara,C., Basvi,P.T., Bapiro,T.E., Sayi,J. and Hasler,J.A. (2004). Frequency of -163 C>A and 63 C>G single nucleotide polymorphism of cytochrome P450 1A2 in two African populations. *Clin. Chem. Lab Med.* **42,** 939-941.

Desmeules,J., Gascon,M.P., Dayer,P. and Magistris,M. (1991). Impact of environmental and genetic factors on codeine analgesia. *Eur. J. Clin. Pharmacol.* **41,** 23-26.

Di Rienzo,A. and Hudson,R.R. (2005). An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* **21,** 596-601.

Eaton,D.L., Gallagher,E.P., Bammler,T.K. and Kunze,K.L. (1995). Role of cytochrome P4501A2 in chemical carcinogenesis: implications for human variability in expression and enzyme activity. *Pharmacogenetics.* **5,** 259-274.

Einarson,T.R. (1993). Drug-related hospital admissions. *Ann. Pharmacother.* **27,** 832-840.

Etheridge,S.L., Spencer,G.J., Heath,D.J. and Genever,P.G. (2004). Expression profiling and functional analysis of wnt signaling mechanisms in mesenchymal stem cells. *Stem Cells.* **22,** 849-860.

Evans,W.E. and McLeod,H.L. (2003). Pharmacogenomics--drug disposition, drug targets, and side effects. *N. Engl. J. Med.* **348,** 538-549.

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

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

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

Excoffier,L., Laval,G. and Balding,D. (2003). Gametic phase estimation over large genomic regions using an adaptive window approach. *Hum. Genomics.* **1,** 7-19.

Facciola,G., Hidestrand,M., von Bahr,C. and Tybring,G. (2001). Cytochrome P450 isoforms involved in melatonin metabolism in human liver microsomes. *Eur. J. Clin. Pharmacol.* **56,** 881-888.

Fang,J., Coutts,R.T., McKenna,K.F. and Baker,G.B. (1998). Elucidation of individual cytochrome P450 enzymes involved in the metabolism of clozapine. *Naunyn Schmiedebergs Arch. Pharmacol.* **358,** 592-599.

Farrall,M. and Weeks,D.E. (1998). Mutational mechanisms for generating microsatellite allele-frequency distributions: an analysis of 4,558 markers. *Am. J. Hum. Genet.* **62,** 1260-1262.

Federal Democratic Republic of Ethiopia Office of Population and Housing Census Commission Central Statistical Authority (1999) The 1994 Population and Housing Census for Ethiopia. Results at Country Level. Volume 2 Analytical Report. In: Central Statistical Authority, Addis Ababa.

Federal Democratic Republic of Ethiopia Ministry of Health (2004) Malaria. Diagnosis and Treatment Guidelines for Health Workers in Ethiopia. In: Addis Ababa.

Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. In: 20 pp 406-416

Florez,J.C., Jablonski,K.A., Bayley,N., Pollin,T.I., de Bakker,P.I., Shuldiner,A.R., Knowler,W.C., Nathan,D.M. and Altshuler,D. (2006). TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N. Engl. J. Med.* **355,** 241-250.

Fontaine,F., de Sousa,G., Burcham,P.C., Duchene,P. and Rahmani,R. (2000). Role of cytochrome P450 3A in the metabolism of mefloquine in human and animal hepatocytes. *Life Sci.* **66,** 2193-2212.

Forster,P. and Matsumura,S. (2005). Evolution. Did early humans go north or south? *Science.* **308,** 965-966.

Freathy,R.M., Weedon,M.N., Bennett,A., Hypponen,E., Relton,C.L., Knight,B., Shields,B., Parnell,K.S., Groves,C.J., Ring,S.M., Pembrey,M.E., Ben Shlomo,Y., Strachan,D.P., Power,C., Jarvelin,M.R., McCarthy,M.I., Davey,S.G., Hattersley,A.T. and Frayling,T.M. (2007). Type 2 diabetes TCF7L2 risk genotypes alter birth weight: a study of 24,053 individuals. *Am. J. Hum. Genet.* **80,** 1150-1161.

Freeman,D. and Pankhurst,A. (2003) Peripheral People. The Excluded Minorities of Ethiopia. In: C. Hurst and Co. Ltd, United Kingdom.

Fu,Y.X. and Li,W.H. (1993). Statistical tests of neutrality of mutations. *Genetics.* **133,** 693-709.

Fuhr,U., Rost,K.L., Engelhardt,R., Sachs,M., Liermann,D., Belloc,C., Beaune,P., Janezic,S., Grant,D., Meyer,U.A. and Staib,A.H. (1996). Evaluation of caffeine as a test drug for CYP1A2, NAT2 and CYP2E1 phenotyping in man by in vivo versus in vitro correlations. *Pharmacogenetics.* **6,** 159-176.

Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M., Liu-Cordero,S.N., Rotimi,C., Adeyemo,A., Cooper,R., Ward,R., Lander,E.S., Daly,M.J. and Altshuler,D. (2002). The structure of haplotype blocks in the human genome. *Science.* **296,** 2225-2229.

Gallagher,E.P., Wienkers,L.C., Stapleton,P.L., Kunze,K.L. and Eaton,D.L. (1994). Role of human microsomal and human complementary DNA-expressed cytochromes P4501A2 and P4503A4 in the bioactivation of aflatoxin B1. *Cancer Res.* **54,** 101-108.

Gasche,Y., Daali,Y., Fathi,M., Chiappe,A., Cottini,S., Dayer,P. and Desmeules,J. (2004). Codeine intoxication associated with ultrarapid CYP2D6 metabolism. *N. Engl. J. Med.* **351,** 2827-2831.

Ghotbi,R., Christensen,M., Roh,H.K., Ingelman-Sundberg,M., Aklillu,E. and Bertilsson,L. (2007). Comparisons of CYP1A2 genetic polymorphisms, enzyme activity and the genotype-phenotype relationship in Swedes and Koreans. *Eur. J. Clin. Pharmacol.* **63,** 537-546.

Giao,P.T. and de Vries,P.J. (2001). Pharmacokinetic interactions of antimalarial agents. *Clin. Pharmacokinet.* **40,** 343-373.

Gloyn,A.L. and McCarthy,M.I. (2001). The genetics of type 2 diabetes. *Best. Pract. Res. Clin. Endocrinol. Metab.* **15,** 293-308.

Goldstein,D.B., Ruiz,L.A., Cavalli-Sforza,L.L. and Feldman,M.W. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 6723-6727.

Goldstone,J.V., Goldstone,H.M., Morrison,A.M., Tarrant,A., Kern,S.E., Woodin,B.R. and Stegeman,J.J. (2007). Cytochrome P450 1 genes in early deuterostomes (tunicates and sea urchins) and vertebrates (chicken and frog): origin and diversification of the CYP1 gene family. *Mol. Biol. Evol.* **24,** 2619-2631.

Gonzalez,F.J. and Gelboin,H.V. (1994). Role of human cytochromes P450 in the metabolic activation of chemical carcinogens and toxins. *Drug Metab Rev.* **26,** 165-183.

Goudet,J., Raymond,M., de Meeus,T. and Rousset,F. (1996). Testing differentiation in diploid populations. *Genetics.* **144,** 1933-1940.

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

Grant,S.F., Thorleifsson,G., Reynisdottir,I., Benediktsson,R., Manolescu,A., Sainz,J., Helgason,A., Stefansson,H., Emilsson,V., Helgadottir,A., Styrkarsdottir,U., Magnusson,K.P., Walters,G.B., Palsdottir,E., Jonsdottir,T., Gudmundsdottir,T., Gylfason,A., Saemundsdottir,J., Wilensky,R.L., Reilly,M.P., Rader,D.J., Bagger,Y., Christiansen,C., Gudnason,V., Sigurdsson,G., Thorsteinsdottir,U., Gulcher,J.R., Kong,A. and Stefansson,K. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38,** 320-323.

Gu,L., Gonzalez,F.J., Kalow,W. and Tang,B.K. (1992). Biotransformation of caffeine, paraxanthine, theobromine and theophylline by cDNA-expressed human CYP1A2 and CYP2E1. *Pharmacogenetics.* **2,** 73-77.

Gunes,A. and Dahl,M.L. (2008). Variation in CYP1A2 activity and its clinical implications: influence of environmental factors and genetic polymorphisms. *Pharmacogenomics.* **9,** 625-637.

Guo,S.W. and Thompson,E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics.* **48,** 361-372.

Hamdy,S.I., Hiratsuka,M., Narahara,K., Endo,N., El Enany,M., Moursi,N., Ahmed,M.S. and Mizugaki,M. (2003). Genotyping of four genetic polymorphisms in the CYP1A2 gene in the Egyptian population. *Br. J. Clin. Pharmacol.* **55,** 321-324.

Han,X.M., Ou-Yang,D.S., Lu,P.X., Jiang,C.H., Shu,Y., Chen,X.P., Tan,Z.R. and Zhou,H.H. (2001). Plasma caffeine metabolite ratio (17X/137X) in vivo associated with G-2964A and C734A polymorphisms of human CYP1A2. *Pharmacogenetics.* **11,** 429-435.

Han,X.M., Ouyang,D.S., Chen,X.P., Shu,Y., Jiang,C.H., Tan,Z.R. and Zhou,H.H. (2002). Inducibility of CYP1A2 by omeprazole in vivo related to the genetic polymorphism of CYP1A2. *Br. J. Clin. Pharmacol.* **54,** 540-543.

Handley,L.J., Manica,A., Goudet,J. and Balloux,F. (2007). Going the distance: human population genetics in a clinal world. *Trends Genet.* **23,** 432-439.

Hardy,O.J. and Vekemans,X. (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity.* **83 (Pt 2),** 145-154.

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

Harris,M.I., Klein,R., Cowie,C.C., Rowland,M. and Byrd-Holt,D.D. (1998). Is the risk of diabetic retinopathy greater in non-Hispanic blacks and Mexican Americans than in non-Hispanic whites with type 2 diabetes? A U.S. population study. *Diabetes Care.* **21,** 1230-1235.

Hartter,S., Ursing,C., Morita,S., Tybring,G., von Bahr,C., Christensen,M., Rojdmark,S. and Bertilsson,L. (2001). Orally given melatonin may serve as a probe drug for cytochrome P450 1A2 activity in vivo: a pilot study. *Clin. Pharmacol. Ther.* **70,** 10-16.

Hayashi,T., Iwamoto,Y., Kaku,K., Hirose,H. and Maeda,S. (2007). Replication study for the association of TCF7L2 with susceptibility to type 2 diabetes in a Japanese population. *Diabetologia.* **50,** 980-984.

Heilmann,L.J., Sheen,Y.Y., Bigelow,S.W. and Nebert,D.W. (1988). Trout P450IA1: cDNA and deduced protein sequence, expression in liver, and evolutionary significance. *DNA.* **7,** 379-387.

Helgason,A., Palsson,S., Thorleifsson,G., Grant,S.F., Emilsson,V., Gunnarsdottir,S., Adeyemo,A., Chen,Y., Chen,G., Reynisdottir,I., Benediktsson,R., Hinney,A., Hansen,T., Andersen,G., Borch-Johnsen,K., Jorgensen,T., Schafer,H., Faruque,M., Doumatey,A., Zhou,J., Wilensky,R.L., Reilly,M.P., Rader,D.J., Bagger,Y., Christiansen,C., Sigurdsson,G., Hebebrand,J., Pedersen,O., Thorsteinsdottir,U., Gulcher,J.R., Kong,A., Rotimi,C. and Stefansson,K. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39,** 218-225.

Henze P.B. (2000) Layers of Time: A History of Ethiopia. In: Palgrave,

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

Horikoshi,M., Hara,K., Ito,C., Nagai,R., Froguel,P. and Kadowaki,T. (2007). A genetic variation of the transcription factor 7-like 2 gene is associated with risk of type 2 diabetes in the Japanese population. *Diabetologia.* **50,** 747-751.

Huang,J.D., Guo,W.C., Lai,M.D., Guo,Y.L. and Lambert,G.H. (1999). Detection of a novel cytochrome P-450 1A2 polymorphism (F21L) in Chinese. *Drug Metab Dispos.* **27,** 98-101.

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

Hughes,A.L., Packer,B., Welch,R., Bergen,A.W., Chanock,S.J. and Yeager,M. (2003). Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 15754-15757.

Hughes,A.L., Packer,B., Welch,R., Bergen,A.W., Chanock,S.J. and Yeager,M. (2005). Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding Loci. *Genetics.* **170,** 1181-1187.

Humphries,S.E., Gable,D., Cooper,J.A., Ireland,H., Stephens,J.W., Hurel,S.J., Li,K.W., Palmen,J., Miller,M.A., Cappuccio,F.P., Elkeles,R., Godsland,I., Miller,G.J. and Talmud,P.J. (2006). Common variants in the TCF7L2 gene and predisposition to type 2 diabetes in UK European Whites, Indian Asians and Afro-Caribbean men and women. *J. Mol. Med.* **84,** 1005-1014.

Ikeya,K., Jaiswal,A.K., Owens,R.A., Jones,J.E., Nebert,D.W. and Kimura,S. (1989). Human CYP1A2: sequence, gene structure, comparison with the mouse and rat orthologous gene, and differences in liver 1A2 mRNA expression. *Mol. Endocrinol.* **3,** 1399-1408.

Ingelman-Sundberg,M. (2001). Pharmacogenetics: an opportunity for a safer and more efficient pharmacotherapy. *J. Intern. Med.* **250,** 186-200.

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

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

Jaiswal,A.K., Nebert,D.W., McBride,O.W. and Gonzalez,F.J. (1987). Human P(3)450: cDNA and complete protein sequence, repetitive Alu sequences in the 3-prime nontranslated region, and localization of gene to chromosome 15. *J. Exp. Path.* **3,** 1-17.

Jakobsson,M., Scholz,S.W., Scheet,P., Gibbs,J.R., VanLiere,J.M., Fung,H.C., Szpiech,Z.A., Degnan,J.H., Wang,K., Guerreiro,R., Bras,J.M., Schymick,J.C., Hernandez,D.G., Traynor,B.J., Simon-Sanchez,J., Matarin,M., Britton,A., van de,L.J., Rafferty,I., Bucan,M., Cann,H.M., Hardy,J.A., Rosenberg,N.A. and Singleton,A.B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature.* **451,** 998-1003.

Jiang,Z., Dalton,T.P., Jin,L., Wang,B., Tsuneoka,Y., Shertzer,H.G., Deka,R. and Nebert,D.W. (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,** 196-206.

Jiang,Z., Dragin,N., Jorge-Nebert,L.F., Martin,M.V., Guengerich,F.P., Aklillu,E., Ingelman-Sundberg,M., Hammons,G.J., Lyn-Cook,B.D., Kadlubar,F.F., Saldana,S.N., Sorter,M., Vinks,A.A., Nassr,N., von Richter,O., Jin,L. and Nebert,D.W. (2006). Search for an association between the human CYP1A2 genotype and CYP1A2 metabolic phenotype. *Pharmacogenet. Genomics.* **16,** 359-367.

Jobling,M.A., Hurles,M.E. and Tyler-Smith,C. (2004) Human Evolutionary Genetics. Origins, Peoples and Disease. In: Garland Science, New York.

Johansson,I., Lundqvist,E., Bertilsson,L., Dahl,M.L., Sjoqvist,F. and Ingelman-Sundberg,M. (1993). Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 11825-11829.

Johansson,I., Oscarson,M., Yue,Q.Y., Bertilsson,L., Sjoqvist,F. and Ingelman-Sundberg,M. (1994). Genetic analysis of the Chinese cytochrome P4502D locus: characterization of variant CYP2D6 genes present in subjects with diminished capacity for debrisoquine hydroxylation. *Mol. Pharmacol.* **46,** 452-459.

Johnson,J.A. (2003). Pharmacogenetics: potential for individualized drug therapy through genetics. *Trends Genet.* **19,** 660-666.

Kalow,W. and Tang,B.K. (1993). The use of caffeine for enzyme assays: a critical appraisal. *Clin. Pharmacol. Ther.* **53,** 503-514.

Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005). LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics.* **21,** 2814-2820.

Kawamoto,K., Lobach,D.F., Willard,H.F. and Ginsburg,G.S. (2009). A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. *BMC. Med. Inform. Decis. Mak.* **9,** 17.

Ke,Y., Su,B., Song,X., Lu,D., Chen,L., Li,H., Qi,C., Marzuki,S., Deka,R., Underhill,P., Xiao,C., Shriver,M., Lell,J., Wallace,D., Wells,R.S., Seielstad,M., Oefner,P., Zhu,D., Jin,J., Huang,W., Chakraborty,R., Chen,Z. and Jin,L. (2001). African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science.* **292,** 1151-1153.

Kendler,K.S. and Prescott,C.A. (1999). Caffeine intake, tolerance, and withdrawal in women: a population-based twin study. *Am. J. Psychiatry.* **156,** 223-228.

Kim,K.A., Park,J.Y., Lee,J.S. and Lim,S. (2003). Cytochrome P450 2C8 and CYP3A4/5 are involved in chloroquine metabolism in human liver microsomes. *Arch. Pharm. Res.* **26,** 631-637.

Kimura,M. (1983). Rare variant alleles in the light of the neutral theory. *Mol. Biol. Evol.* **1,** 84-93.

Kirchheiner,J. and Seeringer,A. (2007). Clinical implications of pharmacogenetics of cytochrome P450 drug metabolizing enzymes. *Biochim. Biophys. Acta.* **1770,** 489-494.

Korinek,V., Barker,N., Moerer,P., van Donselaar,E., Huls,G., Peters,P.J. and Clevers,H. (1998). Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nat. Genet.* **19,** 379-383.

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

Leak,T.S., Keene,K.L., Langefeld,C.D., Gallagher,C.J., Mychaleckyj,J.C., Freedman,B.I., Bowden,D.W., Rich,S.S. and Sale,M.M. (2007). Association of the proprotein convertase subtilisin/kexin-type 2 (PCSK2) gene with type 2 diabetes in an African American population. *Mol. Genet. Metab.* **92,** 145-150.

Lee,A.C., Kamalam,A., Adams,S.M. and Jobling,M.A. (2004). Molecular evidence for absence of Y-linkage of the Hairy Ears trait. *Eur. J. Hum. Genet.* **12,** 1077-1079.

Lee,C.R., Goldstein,J.A. and Pieper,J.A. (2002). Cytochrome P450 2C9 polymorphisms: a comprehensive review of the in-vitro and human data. *Pharmacogenetics.* **12,** 251-263.

Lehman,D.M., Hunt,K.J., Leach,R.J., Hamlington,J., Arya,R., Abboud,H.E., Duggirala,R., Blangero,J., Goring,H.H. and Stern,M.P. (2007). Haplotypes of transcription factor 7-like 2 (TCF7L2) gene and its upstream region are associated with type 2 diabetes and age of onset in Mexican Americans. *Diabetes.* **56,** 389-393.

Lewis I.M. (1998) Peoples of the Horn of Africa: Somali, Afar, and Saho. In: Inc. edition Red Sea Press, Lawrenceville, NJ.

Lewontin,R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics.* **49,** 49-67.

Li,X.Q., Bjorkman,A., Andersson,T.B., Gustafsson,L.L. and Masimirembwa,C.M. (2003). Identification of human cytochrome P(450)s that metabolise anti-parasitic drugs and predictions of in vivo drug hepatic clearance from in vitro data. *Eur. J. Clin. Pharmacol.* **59,** 429-442.

Liang,H.C., Li,H., McKinnon,R.A., Duffy,J.J., Potter,S.S., Puga,A. and Nebert,D.W. (1996). Cyp1a2(-/-) null mutant mice develop normally but show deficient drug metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 1671-1676.

Manica,A., Amos,W., Balloux,F. and Hanihara,T. (2007). The effect of ancient population bottlenecks on human phenotypic variation. *Nature.* **448,** 346-348.

Masimirembwa,C.M. and Hasler,J.A. (1997). Genetic polymorphism of drug metabolising enzymes in African populations: implications for the use of neuroleptics and antidepressants. *Brain Res. Bull.* **44,** 561-571.

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

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

McKusick,V.A. and Francomano,C.A. (1997) Mendelian Inheritance in Man: a catalog of Human Genes and Genetic Disorders. In: Johns Hopkins Univ. Press, Baltimore.

Miljkovic-Gacic,I., Ferrell,R.E., Patrick,A.L., Kammerer,C.M. and Bunker,C.H. (2005). Estimates of African, European and Native American ancestry in Afro-Caribbean men on the island of Tobago. *Hum. Hered.* **60,** 129-133.

Mirghani,R.A., Hellgren,U., Bertilsson,L., Gustafsson,L.L. and Ericsson,O. (2003). Metabolism and elimination of quinine in healthy volunteers. *Eur. J. Clin. Pharmacol.* **59,** 423-427.

Mithen,S. (2007). Did farming arise from a misapplication of social intelligence? *Philos. Trans. R. Soc. Lond B Biol. Sci.* **362,** 705-718.

Miyata,T., Miyazawa,S. and Yasunaga,T. (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12,** 219-236.

Murayama,N., Soyama,A., Saito,Y., Nakajima,Y., Komamura,K., Ueno,K., Kamakura,S., Kitakaze,M., Kimura,H., Goto,Y., Saitoh,O., Katoh,M., Ohnuma,T., Kawai,M., Sugai,K., Ohtsuki,T., Suzuki,C., Minami,N., Ozawa,S. and Sawada,J. (2004). Six novel nonsynonymous CYP1A2 gene polymorphisms: catalytic activities of the naturally occurring variant enzymes. *J. Pharmacol. Exp. Ther.* **308,** 300-306.

Nakajima,M., Yokoi,T., Mizutani,M., Shin,S., Kadlubar,F.F. and Kamataki,T. (1994). Phenotyping of CYP1A2 in Japanese population by analysis of caffeine urinary metabolites: absence of mutation prescribing the phenotype in the CYP1A2 gene. *Cancer Epidemiol. Biomarkers Prev.* **3,** 413-421.

Nakajima,M., Yokoi,T., Mizutani,M., Kinoshita,M., Funayama,M. and Kamataki,T. (1999). Genetic polymorphism in the 5'-flanking region of human CYP1A2 gene: effect on the CYP1A2 inducibility in humans. *J. Biochem. (Tokyo).* **125,** 803-808.

Nebert,D.W. and Russell,D.W. (2002). Clinical importance of the cytochromes P450. *Lancet.* **360,** 1155-1162.

NEEL,J.V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am. J. Hum. Genet.* **14,** 353-362.

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

Nelson,D.R., Koymans,L., Kamataki,T., Stegeman,J.J., Feyereisen,R., Waxman,D.J., Waterman,M.R., Gotoh,O., Coon,M.J., Estabrook,R.W., Gunsalus,I.C. and Nebert,D.W. (1996). P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics.* **6,** 1-42.

Nelson,D.R. (2009). The cytochrome p450 homepage. *Hum. Genomics.* **4,** 59-65.

Niu,T., Qin,Z.S., Xu,X. and Liu,J.S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70,** 157-169.

Nordmark,A., Lundgren,S., Ask,B., Granath,F. and Rane,A. (2002). The effect of the CYP1A2 *1F mutation on CYP1A2 inducibility in pregnant women. *Br. J. Clin. Pharmacol.* **54,** 504-510.

Ohta,T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature.* **246,** 96-98.

Ohta,T. (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Popul. Biol.* **10,** 254-275.

Ohta,T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16134-16137.

Oldroyd,J., Banerjee,M., Heald,A. and Cruickshank,K. (2005). Diabetes and ethnic minorities. *Postgrad. Med. J.* **81,** 486-490.

OMURA,T. and SATO,R. (1964). THE CARBON MONOXIDE-BINDING PIGMENT OF LIVER MICROSOMES. I. EVIDENCE FOR ITS HEMOPROTEIN NATURE. *J. Biol. Chem.* **239,** 2370-2378.

Orlando,R., Piccoli,P., De Martin,S., Padrini,R., Floreani,M. and Palatini,P. (2004). Cytochrome P450 1A2 is a major determinant of lidocaine metabolism in vivo: effects of liver function. *Clin. Pharmacol. Ther.* **75,** 80-88.

Oscarson,M., Hidestrand,M., Johansson,I. and Ingelman-Sundberg,M. (1997). A combination of mutations in the CYP2D6*17 (CYP2D6Z) allele causes alterations in enzyme function. *Mol. Pharmacol.* **52,** 1034-1040.

Pankhurst P. (2001) *The Ethiopians: A History*. In: Wiley-Blackwell,

Papadopoulou,S. and Edlund,H. (2005). Attenuated Wnt signaling perturbs pancreatic growth but not pancreatic function. *Diabetes.* **54,** 2844-2851.

Paradies,Y.C., Montoya,M.J. and Fullerton,S.M. (2007). Racialized genetics and the study of complex diseases: the thrifty genotype revisited. *Perspect. Biol. Med.* **50,** 203-227.

Parkinson,A., Mudra,D.R., Johnson,C., Dwyer,A. and Carroll,K.M. (2004). The effects of gender, age, ethnicity, and liver cirrhosis on cytochrome P450 enzyme activity in human liver microsomes and inducibility in cultured human hepatocytes. *Toxicol. Appl. Pharmacol.* **199,** 193-209.

Pavanello,S., Pulliero,A., Lupi,S., Gregorio,P. and Clonfero,E. (2005). Influence of the genetic polymorphism in the 5'-noncoding region of the CYP1A2 gene on CYP1A2 phenotype and urinary mutagenicity in smokers. *Mutat. Res.* **587,** 59-66.

Phillips,K.A., Veenstra,D.L., Oren,E., Lee,J.K. and Sadee,W. (2001). Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. *JAMA.* **286,** 2270-2279.

Pickwell,G.V., Shih,H. and Quattrochi,L.C. (2003). Interaction of upstream stimulatory factor proteins with an E-box located within the human CYP1A2 5'-flanking gene contributes to basal transcriptional gene activation. *Biochem. Pharmacol.* **65,** 1087-1096.

Pirmohamed,M. and Park,B.K. (2003). Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology.* **192,** 23-32.

Pirmohamed,M., James,S., Meakin,S., Green,C., Scott,A.K., Walley,T.J., Farrar,K., Park,B.K. and Breckenridge,A.M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ.* **329,** 15-19.

Porter,T.D. and Coon,M.J. (1991). Cytochrome P-450. Multiplicity of isoforms, substrates, and catalytic and regulatory mechanisms. *J. Biol. Chem.* **266,** 13469-13472.

Pritchard,J.K., Stephens,M., Rosenberg,N.A. and Donnelly,P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* **67,** 170-181.

Projean,D., Baune,B., Farinotti,R., Flinois,J.P., Beaune,P., Taburet,A.M. and Ducharme,J. (2003). In vitro metabolism of chloroquine: identification of CYP2C8, CYP3A4, and CYP2D6 as the main isoforms catalyzing N-desethylchloroquine formation. *Drug Metab Dispos.* **31,** 748-754.

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

Przeworski,M. (2002). The signature of positive selection at randomly chosen loci. *Genetics.* **160,** 1179-1189.

Quattrochi,L.C., Vu,T. and Tukey,R.H. (1994). The human CYP1A2 gene and induction by 3-methylcholanthrene. A region of DNA that supports AH-receptor binding and promoter-specific induction. *J. Biol. Chem.* **269,** 6949-6954.

Quattrochi,L.C., Shih,H. and Pickwell,G.V. (1998). Induction of the human CYP1A2 enhancer by phorbol ester. *Arch. Biochem. Biophys.* **350,** 41-48.

Ramachandran,S., Deshpande,O., Roseman,C.C., Rosenberg,N.A., Feldman,M.W. and Cavalli-Sforza,L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15942-15947.

Rasmussen,B.B., Brix,T.H., Kyvik,K.O. and Brosen,K. (2002). The interindividual differences in the 3-demthylation of caffeine alias CYP1A2 is determined by both genetic and environmental factors. *Pharmacogenetics.* **12,** 473-478.

Reed,F.A. and Tishkoff,S.A. (2006). African human diversity, origins and migrations. *Curr. Opin. Genet. Dev.* **16,** 597-605.

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

Rezen,T., Contreras,J.A. and Rozman,D. (2007). Functional genomics approaches to studies of the cytochrome p450 superfamily. *Drug Metab Rev.* **39,** 389-399.

Riste,L., Khan,F. and Cruickshank,K. (2001). High prevalence of type 2 diabetes in all ethnic groups, including Europeans, in a British inner city: relative poverty, history, inactivity, or 21st century Europe? *Diabetes Care.* **24,** 1377-1383.

Rosenberg,N.A., Pritchard,J.K., Weber,J.L., Cann,H.M., Kidd,K.K., Zhivotovsky,L.A. and Feldman,M.W. (2002). Genetic structure of human populations. *Science.* **298,** 2381-2385.

Rosenberg,N.A., Mahajan,S., Gonzalez-Quevedo,C., Blum,M.G., Nino-Rosales,L., Ninis,V., Das,P., Hegde,M., Molinari,L., Zapata,G., Weber,J.L., Belmont,J.W. and Patel,P.I. (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS. Genet.* **2,** e215.

Ross,S.E., Hemati,N., Longo,K.A., Bennett,C.N., Lucas,P.C., Erickson,R.L. and MacDougald,O.A. (2000). Inhibition of adipogenesis by Wnt signaling. *Science.* **289,** 950-953.

Rotimi,C.N., Chen,G., Adeyemo,A.A., Furbert-Harris,P., Parish-Gause,D., Zhou,J., Berg,K., Adegoke,O., Amoah,A., Owusu,S., Acheampong,J., Agyenim-Boateng,K., Eghan,B.A., Jr., Oli,J., Okafor,G., Ofoegbu,E., Osotimehin,B., Abbiyesuku,F., Johnson,T., Rufus,T., Fasanmade,O., Kittles,R., Daniel,H., Chen,Y., Dunston,G. and Collins,F.S. (2004). A genome-wide search for type 2 diabetes susceptibility genes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. *Diabetes.* **53,** 838-841.

Rotimi,C.N., Chen,G., Adeyemo,A.A., Jones,L.S., Agyenim-Boateng,K., Eghan,B.A., Jr., Zhou,J., Doumatey,A., Lashley,K., Huang,H., Fasanmade,O., Akinsola,F.B., Ezepue,F., Amoah,A., Akafo,S., Chen,Y., Oli,J. and Johnson,T. (2006). Genomewide scan and fine mapping of quantitative trait loci for intraocular pressure on 5q and 14q in West Africans. *Invest Ophthalmol. Vis. Sci.* **47,** 3262-3267.

Rousset,F. and Raymond,M. (1995). Testing heterozygote excess and deficiency. *Genetics.* **140,** 1413-1419.

Sachse,C., Brockmoller,J., Bauer,S. and Roots,I. (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,** 445-449.

Sachse,C., Bhambra,U., Smith,G., Lightfoot,T.J., Barrett,J.H., Scollay,J., Garner,R.C., Boobis,A.R., Wolf,C.R. and Gooderham,N.J. (2003). Polymorphisms in the cytochrome P450 CYP1A2 gene (CYP1A2) in colorectal cancer patients and controls: allele frequencies, linkage disequilibrium and influence on caffeine metabolism. *Br. J. Clin. Pharmacol.* **55,** 68-76.

Saito,Y., Hanioka,N., Maekawa,K., Isobe,T., Tsuneto,Y., Nakamura,R., Soyama,A., Ozawa,S., Tanaka-Kagawa,T., Jinno,H., Narimatsu,S. and Sawada,J. (2005). Functional analysis of three CYP1A2 variants found in a Japanese population. *Drug Metab Dispos.* **33,** 1905-1910.

Sakuyama,K., Sasaki,T., Ujiie,S., Obata,K., Mizugaki,M., Ishikawa,M. and Hiratsuka,M. (2008). Functional characterization of 17 CYP2D6 allelic variants (CYP2D6.2, 10, 14A-B, 18, 27, 36, 39, 47-51, 53-55, and 57). *Drug Metab Dispos.* **36,** 2460-2467.

Sansen,S., Yano,J.K., Reynald,R.L., Schoch,G.A., Griffin,K.J., Stout,C.D. and Johnson,E.F. (2007). Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *J. Biol. Chem.* **282,** 14348-14355.

Sarkar,M.A., Hunt,C., Guzelian,P.S. and Karnes,H.T. (1992). Characterization of human liver cytochromes P-450 involved in theophylline metabolism. *Drug Metab Dispos.* **20,** 31-37.

Saxena,R., Gianniny,L., Burtt,N.P., Lyssenko,V., Giuducci,C., Sjogren,M., Florez,J.C., Almgren,P., Isomaa,B., Orho-Melander,M., Lindblad,U., Daly,M.J., Tuomi,T., Hirschhorn,J.N., Ardlie,K.G., Groop,L.C. and Altshuler,D. (2006). Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes.* **55,** 2890-2895.

Scheet,P. and Stephens,M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78,** 629-644.

Schneider, Roessli D. and Excoffier L. (2000) Arlequin: A software for population genetics data analysis. In: Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Schweikl,H., Taylor,J.A., Kitareewan,S., Linko,P., Nagorney,D. and Goldstein,J.A. (1993). Expression of CYP1A1 and CYP1A2 genes in human liver. *Pharmacogenetics.* **3,** 239-249.

Scott,L.J., Bonnycastle,L.L., Willer,C.J., Sprau,A.G., Jackson,A.U., Narisu,N., Duren,W.L., Chines,P.S., Stringham,H.M., Erdos,M.R., Valle,T.T., Tuomilehto,J., Bergman,R.N., Mohlke,K.L., Collins,F.S. and Boehnke,M. (2006). Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. *Diabetes.* **55,** 2649-2653.

Shet,M.S., McPhaul,M., Fisher,C.W., Stallings,N.R. and Estabrook,R.W. (1997). Metabolism of the antiandrogenic drug (Flutamide) by human CYP1A2. *Drug Metab Dispos.* **25,** 1298-1303.

Shimada,T., Yamazaki,H., Mimura,M., Inui,Y. and Guengerich,F.P. (1994). Interindividual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians. *J. Pharmacol. Exp. Ther.* **270,** 414-423.

Shimoda,K., Someya,T., Morita,S., Hirokane,G., Yokono,A., Takahashi,S. and Okawa,M. (2002). Lack of impact of CYP1A2 genetic polymorphism (C/A polymorphism at position 734 in intron 1 and G/A polymorphism at position -2964 in the 5'-flanking region of CYP1A2) on the plasma concentration of haloperidol in smoking male Japanese with schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry.* **26,** 261-265.

Shreeve,J. (1994). 'Lucy,' crucial early human ancestor, finally gets a head. *Science.* **264,** 34-35.

Singh,R., Shaw,J. and Zimmet,P. (2004). Epidemiology of childhood type 2 diabetes in the developing world. *Pediatr. Diabetes.* **5,** 154-168.

Sladek,R., Rocheleau,G., Rung,J., Dina,C., Shen,L., Serre,D., Boutin,P., Vincent,D., Belisle,A., Hadjadj,S., Balkau,B., Heude,B., Charpentier,G., Hudson,T.J., Montpetit,A., Pshezhetsky,A.V., Prentki,M., Posner,B.I., Balding,D.J., Meyre,D., Polychronakos,C. and Froguel,P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature.* **445,** 881-885.

Slatkin,M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics.* **139,** 457-462.

Smith,U. (2007). TCF7L2 and type 2 diabetes--we WNT to know. *Diabetologia.* **50,** 5-7.

Solus,J.F., Arietta,B.J., Harris,J.R., Sexton,D.P., Steward,J.Q., McMunn,C., Ihrie,P., Mehall,J.M., Edwards,T.L. and Dawson,E.P. (2004). Genetic variation in eleven phase I drug metabolism genes in an ethnically diverse population. *Pharmacogenomics.* **5,** 895-931.

Soyama,A., Saito,Y., Hanioka,N., Maekawa,K., Komamura,K., Kamakura,S., Kitakaze,M., Tomoike,H., Ueno,K., Goto,Y., Kimura,H., Katoh,M., Sugai,K., Saitoh,O., Kawai,M., Ohnuma,T., Ohtsuki,T., Suzuki,C., Minami,N., Kamatani,N., Ozawa,S. and Sawada,J. (2005). Single nucleotide polymorphisms and haplotypes of CYP1A2 in a Japanese population. *Drug Metab Pharmacokinet.* **20,** 24-33.

Stephens,M., Smith,N.J. and Donnelly,P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978-989.

Stephens,M. and Donnelly,P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73,** 1162-1169.

Stephens,M. and Scheet,P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76,** 449-462.

Stringer,C.B. and Andrews,P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science.* **239,** 1263-1268.

Sugahara,H., Maebara,C., Ohtani,H., Handa,M., Ando,K., Mine,K., Kubo,C., Ieiri,I. and Sawada,Y. (2009). Effect of smoking and CYP2D6 polymorphisms on the extent of fluvoxamine-alprazolam interaction in patients with psychosomatic disease. *Eur. J. Clin. Pharmacol.*

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

Tan,E.K., Chua,E., Fook-Chong,S.M., Teo,Y.Y., Yuen,Y., Tan,L. and Zhao,Y. (2007). Association between caffeine intake and risk of Parkinson's disease among fast and slow metabolizers. *Pharmacogenet. Genomics.* **17,** 1001-1005.

Thomas,M.G., Skorecki,K., Ben Ami,H., Parfitt,T., Bradman,N. and Goldstein,D.B. (1998). Origins of Old Testament priests. *Nature.* **394,** 138-140.

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

Tishkoff,S.A., Dietzsch,E., Speed,W., Pakstis,A.J., Kidd,J.R., Cheung,K., Bonne-Tamir,B., Santachiara-Benerecetti,A.S., Moral,P. and Krings,M. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science.* **271,** 1380-1387.

Tishkoff,S.A. and Verrelli,B.C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* **4,** 293-340.

Tishkoff,S.A. and Kidd,K.K. (2004). Implications of biogeography of human populations for 'race' and medicine. *Nat. Genet.* **36,** S21-S27.

Tishkoff,S.A., Reed,F.A., Friedlaender,F.R., Ehret,C., Ranciaro,A., Froment,A., Hirbo,J.B., Awomoyi,A.A., Bodo,J.M., Doumbo,O., Ibrahim,M., Juma,A.T., Kotze,M.J., Lema,G., Moore,J.H., Mortensen,H., Nyambo,T.B., Omar,S.A., Powell,K., Pretorius,G.S., Smith,M.W., Thera,M.A., Wambebe,C., Weber,J.L. and Williams,S.M. (2009). The genetic structure and history of Africans and African Americans. *Science.* **324,** 1035-1044.

Tremblay,M. and Vezina,H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.* **66,** 651-658.

Van Ess,P.J., Pedersen,W.A., Culmsee,C., Mattson,M.P. and Blouin,R.A. (2002). Elevated hepatic and depressed renal cytochrome P450 activity in the Tg2576 transgenic mouse model of Alzheimer's disease. *J. Neurochem.* **80,** 571-578.

Vestal T.M. (1999) Ethiopia: A post-Cold War African state. In: Praeger Publishers,

Wall,J.D. (1999). Recombination and the power of statistical tests of neutrality. *Genetical Research.* **74,** 65-79.

Wallace,D.C., Brown,M.D. and Lott,M.T. (1999). Mitochondrial DNA variation in human evolution and disease. *Gene.* **238,** 211-230.

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

Weedon,M.N. (2007). The importance of TCF7L2. *Diabet. Med.* **24,** 1062-1066.

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

Welfare,M.R., Aitkin,M., Bassendine,M.F. and Daly,A.K. (1999). Detailed modelling of caffeine metabolism and examination of the CYP1A2 gene: lack of a polymorphism in CYP1A2 in Caucasians. *Pharmacogenetics.* **9,** 367-375.

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

Wiegand,S., Maikowski,U., Blankenstein,O., Biebermann,H., Tarnow,P. and Gruters,A. (2004). Type 2 diabetes and impaired glucose tolerance in European children and adolescents with obesity -- a problem that is no longer restricted to minority groups. *Eur. J. Endocrinol.* **151,** 199-206.

Wild,S., Roglic,G., Green,A., Sicree,R. and King,H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care.* **27,** 1047-1053.

Wilke,R.A., Lin,D.W., Roden,D.M., Watkins,P.B., Flockhart,D., Zineh,I., Giacomini,K.M. and Krauss,R.M. (2007). Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat. Rev. Drug Discov.* **6,** 904-916.

Wilkinson,G.R. (2005). Drug metabolism and variability among patients in drug response. *N. Engl. J. Med.* **352,** 2211-2221.

Wong,G.K., Yang,Z., Passey,D.A., Kibukawa,M., Paddock,M., Liu,C.R., Bolund,L. and Yu,J. (2003). A population threshold for functional polymorphisms. *Genome Res.* **13,** 1873-1879.

Wong,N.A. and Pignatelli,M. (2002). Beta-catenin--a linchpin in colorectal carcinogenesis? *Am. J. Pathol.* **160,** 389-401.

Xu,C.F., Lewis,K., Cantone,K.L., Khan,P., Donnelly,C., White,N., Crocker,N., Boyd,P.R., Zaykin,D.V. and Purvis,I.J. (2002). Effectiveness of computational methods in haplotype prediction. *Hum. Genet.* **110,** 148-156.

Yi,F., Brubaker,P.L. and Jin,T. (2005). TCF-4 mediates cell type-specific regulation of proglucagon gene expression by beta-catenin and glycogen synthase kinase-3beta. *J. Biol. Chem.* **280,** 1457-1464.

Zaccaro,C., Sweitzer,S., Pipino,S., Gorman,N., Sinclair,P.R., Sinclair,J.F., Nebert,D.W. and De Matteis,F. (2001). Role of cytochrome P450 1A2 in bilirubin degradation Studies in Cyp1a2 (-/-) mutant mice. *Biochem. Pharmacol.* **61,** 843-849.

Zeggini,E. and McCarthy,M.I. (2007). Identifying susceptibility variants for type 2 diabetes. *Methods Mol. Biol.* **376,** 235-250.

Zeng,Z., Andrew,N.W., Arison,B.H., Luffer-Atlas,D. and Wang,R.W. (1998). Identification of cytochrome P4503A4 as the major enzyme responsible for the metabolism of ivermectin by human liver microsomes. *Xenobiotica.* **28,** 313-321.

Zhang,S., Pakstis,A.J., Kidd,K.K. and Zhao,H. (2001). Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am. J. Hum. Genet.* **69,** 906-914.

Zhivotovsky,L.A., Rosenberg,N.A. and Feldman,M.W. (2003). Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72,** 1171-1186.

Zhou,H., Josephy,P.D., Kim,D. and Guengerich,F.P. (2004). Functional characterization of four allelic variants of human cytochrome P450 1A2. *Arch. Biochem. Biophys.* **422,** 23-30.