Molecular diversity and population structure at the *CYP3A5* gene in Africa

Ripudaman Kaur Bains

Submitted for the Doctor of Philosophy degree

2012

Research Department of Genetics, Evolution and Environment,
University College London

Primary Supervisor: Professor Andrés Ruiz-Linares M.D., Ph.D.
Second Supervisor: Professor Elizabeth Shephard Ph.D.

Declaration of authorship

I, Ripudaman Kaur Bains, 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.

Each image that has copyright restrictions, and/or been published, has been reproduced with permission from the holder of the copyright, and/or the author of the original publication.

Abstract

The CYP450 superfamily of enzymes metabolise ~90% of all therapeutic drugs. CYP3A5 is involved in the metabolism of multiple drugs and endogenous compounds. Enzyme expression is highly variable and associated with differential efficacy of therapeutic drugs and risk of adverse drug reactions (ADRs).

Four functionally important *CYP3A5* variants: *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7,* have been identified in broad human population surveys. *CYP3A5*1* produces a functional protein while *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* define variants which reduce enzyme expression. Reduced CYP3A5 expression is associated with ADRs. Conversely elevated *CYP3A5*3* frequencies are observed in non-equatorial populations and have been reported to protect against the onset of salt-sensitive hypertension.

Little is known about *CYP3A5* variability in Africa; a region that has more genetic diversity than the rest of the world combined. The main objectives of this thesis were to characterise intra-African variation in the *CYP3A5* gene; identify likely implications of *CYP3A5* variability on African healthcare; and examine evidence of selection on the gene. Appreciable African frequencies of *CYP3A5*6* (12-33%) and *CYP3A5*7* (3-22%) were identified and are likely to contribute to variable CYP3A5 expression across the continent. *CYP3A5*6* was observed in every genotyped African population; *CYP3A5*7* was observed almost exclusively in Niger-Congo speaking populations. Evidence of positive selection acting on *CYP3A5* was found and coalescent dates of low/non-expresser *CYP3A5* variants indicate that *CYP3A5*3* is likely to have undergone a recent, rapid, increase in frequency in non-African populations.

Re-sequencing of a ~12kb *CYP3A5* region in five Ethiopian populations; and a ~4.5kb region in eight additional African populations, identified additional variants which may cause low/non-expression of CYP3A5.

Considerable intra-African differences in *CYP3A5* allele frequencies and haplotype structure were identified. Intra-African *CYP3A5* variability suggests that there is likely to be differential efficacy of CYP3A5 metabolised drugs, and associated susceptibility to ADRs between individuals and groups across the continent.

Acknowledgements

There are many people whose support and enthusiasm for this project I am very grateful for. Firstly thank you to BBSRC for funding this project. Unquestionably my biggest thank you goes to my principal supervisor Professor Andrés Ruiz-Linares; for his invaluable advice, support, good humour and encouragement, without which I simply would not have been able to complete my Ph.D.

My acknowledgements would not be complete without thanking my second supervisor Professor Elizabeth Shephard, as well as Professor Dallas Swallow, Professor Kevin Fowler and Dr Julia Day; who have all been wonderfully supportive throughout my Ph.D.

Thank you also to Professor Mark Thomas for his help, input, support and good humour throughout this project. Thanks also to Professor Thomas and to Dr Neil Bradman, of the Melford Charitable Trust, for allowing me access to the TCGA sample collection; one of the best in the world. Additional thanks to the Melford Charitable Trust for providing supplementary funding for this project.

I am very grateful to a number of Ph.D. colleagues and friends who have made my four years so enjoyable; in particular Bryony Jones and Barbara Kremeyer, my fellow "Shanghai ladies", for their superb company in Shanghai, without which the trip would not have been anywhere near as much fun. Also thanks to Pawel Zmarz, Sushma Jansari, Mari-Wyn Burley, Anke Liebert, Kate Brown, Marijke Frantsen, Katherine Brown, Claire Peart and Larissa Kogleck who all never failed to make me laugh on a daily basis, and for providing much needed distractions every now and then; to Victor Acuña Alonzo, Pascale Gerbault, Adrian Timpson, Mirna Kovasevic and Yuval Itan for their advice on many aspects of this project. Thank you also to Ranji Arasaretnam, Christopher Plaster, Gurjeet Rajbans, Jane Dempster, Siobhan Cox, Antonia Ford, Judy Savage, Nicolas Montalva, Rosemary Ekong, Ayele Tarekegn, Endashaw Bekele, Laura Horsfall, Krishna Veeramah and Sijia Wang for all of their help, advice and company during my Ph.D.

Thank you to my many friends outside of the lab (many of whom have shared the pain of a labour of love that is the Ph.D. thesis)! In particular, Bobbi Pritt and her amazing editing skills, Rebecca Aggarwal, Alex Ball, Bhavini Patel, Tankut Guney and Risa Mori who all make me laugh every time I see and speak to them. Most of all, I thank Jonathan Viney for always keeping me grounded in both my academic and personal life.

Finally, I thank my family: Har Anoop, Parl and Mohipinder for their support and interest in my research, without which I would not have been able to complete my Ph.D. Most of all, I thank my beloved parents: my mother Ninder and my father the late Sukhwinder Singh Bains (Kikky). They both taught me the most important lesson of my life: to always aim for the sky in the hope of touching a few stars. I dedicate this thesis to them for believing in me.

Ripudaman Kaur Bains, 2012

Table of contents

List of Figures

List of Tables

1 Introduction

This thesis is concerned with the molecular and population genetics of variation in the gene encoding the human drug metabolising enzyme Cytochrome P450 3A5 (CYP3A5) in sub-Saharan Africa; with a special emphasis on Ethiopian populations. CYP3A5 plays an important role in the metabolism of many endogenous and exogenous substrates (Patki et al. 2003) including a wide spectrum of drugs in clinical use (Frohlich et al. 2004; Wong et al. 2004; Mirghani et al. 2006). It is also of particular interest in medical research due to its role in predicting disease pathology and risk (Givens et al. 2003; Zhenhua et al. 2005). Despite its importance, enzyme expression is polymorphic; individuals can express the protein at high concentrations or have low/undetectable levels of enzyme (Kuehl et al. 2001). A number of factors, such as demography and drift, can cause inter-population differences in allele frequencies, however it is becoming increasingly clear that for the *CYP3A5* locus, selection has played an important role (Thompson et al. 2004; Chen et al. 2009).

This introduction reviews the current literature on the impact of human genetic variation on healthcare. The importance of understanding the role of ancestry in healthcare is also addressed; with reference to sub-Saharan African populations. The background on the *CYP3A5* gene is then described in detail along with a review of the evidence for its role in drug metabolism and in predicting disease risk and pathology. Finally an overview of methods used to evaluate selection, and a review of the evidence for selection on the *CYP3A5* gene, is discussed. The specific aims and outline of this thesis are described in detail thereafter.

1.1 Part I: Human genetic variation and healthcare implications

1.1.1 Human genetic variation

The human genome exhibits considerable inter-individual sequence variation between and within population groups. This variation may be influenced by a number of demographic factors including fluctuations in population size, and random mutation. Selection and genetic drift within and between populations can lead to differentiation of sub-sections of individuals from a wider population.

Genomic variation can be both significant and non-significant; depending upon any consequent effect on transcription and translation of particular proteins, and any associated phenotypic effect of these changes. Variation in the genome can range from single base

changes (single nucleotide polymorphisms: SNPs) to large regions of genomic sequence(s) being inserted, deleted (INDELs) or copied (copy-number variants: CNV).

Since the publication of the human genome sequence in 2000 (Venter et al. 2001), millions of SNPs have been identified. Humans have been reported to differ at the single nucleotide level at approximately 1 in every 1000 base pairs (Sachidanandam et al. 2001; Rotimi and Jorde 2010). A principal challenge of the genomics era is to identify functionally important variation within the human genome in order to reconstruct human evolutionary history and to understand the genetic basis of human diseases. Improved understanding of the genetic determinants of human diseases will inevitably aid efforts to treat endemic and emerging global infections (Tishkoff and Verrelli 2003). Much research has focused on identifying medically important genetic variants. Many studies have mapped disease causing genes and variants using methods which exploit the existence of linkage disequilibrium (Goldstein and Weale 2001) where a set of markers have been observed to be associated with a particular phenotype and then mapped to specific chromosomes.

Recently there has been a move towards examining structural variation in the human genome. Structural variation defines genomic alterations that affect DNA structure and arrangement of the genome (Scherer et al. 2007). Structural variation commonly includes INDELs and CNV which have been the focus of recent studies attempting to identify medically important variation (Patrinos and Petersen 2009; Wain et al. 2009), in addition to SNP based analysis (Giacomini et al. 2007).

Genetic variability does not only influence disease phenotype and pathology, it can also influence the efficacy of drug treatment. These studies are now increasingly frequent and have broadened our understanding of clinically relevant genetic variation (Giacomini et al. 2007).

1.1.2 Pharmacogenetics

Pharmacogenetics is the combination of pharmacological and molecular genetics based fields to determine how genetic factors may affect the efficacy and safety of drug treatment (Weinshilboum 2003; Wilke et al. 2007; Johnson 2008). The clinical vision for pharmacogenetics is that genetic information might be used to identify drugs and dosages that have the most beneficial treatment outcome for an individual patient.

The results of drug therapy can vary within and between populations. Whilst many patients respond well to drug treatment, there are individuals who have minimal or no therapeutic response, additionally some patients experience severe adverse drug reactions

which are major contributors to global morbidity and mortality (Weinshilboum 2003; Wilke et al. 2007).

A number of different factors may affect the ability of a patient to respond effectively to drug treatment including sex, age, underlying co-morbidities, stage of disease and concurrent medications. Aside from environmental factors which may influence drug efficacy and safety, genetic factors are also important.

Variation in drug metabolising enzymes has been shown to have conflicting effects on the ability of an individual to metabolise a variety of drugs including warfarin, HIV-1 protease inhibitors and antidepressant medication (Lin et al. 2002; Givens et al. 2003; Pirmohamed and Park 2003; Weinshilboum 2003; Frohlich et al. 2004; Mouly et al. 2005; Lynch and Price 2007; Wilke et al. 2007; Kohlrausch et al. 2008). Once a drug is administered, it is absorbed and distributed to the site of action, where it interacts with targets such as receptors and/or enzymes. Most drugs undergo metabolism before being excreted. Genetic variation may affect absorption, enzyme activity, cellular uptake, and metabolism, resulting in altered drug activity or half-life (Weinshilboum 2003). When compared to the effect of complex and potentially interacting environmental influences, genetic factors that alter the pharmacodynamics of a particular drug may be easier to detect.

The implementation of genotypic guided medicine for individuals is not in widespread clinical use. Although physicians are becoming increasingly aware of clinically relevant genetic polymorphisms; a 2006 study by the Federal Drug Administration reported that ~25% of all prescriptions written in the USA contained pharmacogenetics labelling (Gardiner and Begg 2006). The paucity of affordable and efficient testing methods; and the continuous identification of clinically important genetic variants are factors which have delayed the translation of human genetic information into clinical practice and healthcare administration (Weinshilboum 2003; Constable et al. 2006).

Population-based studies have been invaluable in filling this gap. Individuals of a given population may have underlying genetic similarities which could potentially distinguish them from other populations. The focus on identifying important variants within populations instead of individuals has identified common, medically important, variation (Wojnowski et al. 2004; Dandara et al. 2005; Mirghani et al. 2006; Gebeyehu et al. 2011).

1.1.3 Ancestry and human health

Many genetic studies have identified common variation, i.e. that which is observed at a frequency greater than 5% in all human populations (Rotimi and Jorde 2010). However it is

important to mention that perceptions of population differences are strongly influenced by which human populations have been sampled.

Although common genetic variation has been identified between populations, there are well-established inter-ethnic and inter-population differences in disease prevalence. Examples include higher frequencies of Tay-Sachs disease in Ashkenazi Jewish populations (Myerowitz 2001; Frisch et al. 2004) and high frequencies of blood disorders in populations within malaria endemic regions, such as glucose-6-phosphate dehydrogenase deficiency (Tishkoff et al. 2001) and sickle cell disorder (Mabayoje 1956; Trowell et al. 1957).

However, there are disparities in population-based healthcare which are due to sampling bias. Approximately 90% of genome wide association studies, aiming to identify genetic variation important in disease susceptibility, have been performed in populations with recent European ancestry (hereafter called European populations). Only one major genome wide study on disease causing variants has been performed in individuals with recent African ancestry which examined variants important in malaria causation and phenotype (Rotimi and Jorde 2010).

Large sample and ascertainment biases in studies on human genetic variation do not account for considerable diversity within African populations (Reed and Tishkoff 2006). Many disease-related SNP microarrays are biased towards European populations and so overlook diversity within populations with recent African ancestry (Browning et al. 2010; Schuster et al. 2010).

1.1.4 Sub-Saharan Africans in evolutionary and medical-based research

The importance of including sub-Saharan Africans as study populations within evolutionary and medical based research should not be underestimated. From an evolutionary research perspective, sub-Saharan Africa is essential for reconstructing human evolutionary history and in understanding how demographic factors, such as changes in population size, and long range migration, have influenced diversity within human populations, as has selection (Campbell and Tishkoff 2008).

Sub-Saharan Africa is a region which extends from just below the Sahara desert to South Africa, Figure 1.1. The region has extensive genetic, cultural, linguistic and phenotypic diversity; only a fraction of which has been observed outside of the sub-continent (Tishkoff and Verrelli 2003; Manica et al. 2007; Campbell and Tishkoff 2008). Approximately 30.5% (over 2000) of the world's languages, representing multiple language families, are spoken in the region (http://www.ethnologue.com), see Figure 1.2. Populations on the continent inhabit

diverse environments (Reed and Tishkoff 2006), which have been susceptible to change over thousands of years, simultaneously with human evolution (Scholz et al. 2007) and are likely to have influenced diversity within the region.

Genetic variation within sub-Saharan Africa has been influenced by demographic events over 200,000 years: including short and long range migration, population admixture and fluctuations in population size (Tishkoff and Verrelli 2003; Manica et al. 2007; Campbell and Tishkoff 2008; Campbell and Tishkoff 2010).

Figure 1.1: A political map of sub-Saharan Africa, image has been taken from: (http://www.agricultureinformation.com/mag/wp-content/uploads/2009/04/africa-map-e.jpg).

Figure 1.2: A map showing the distribution of major language families in Africa. Image has been taken from (http://news.sciencemag.org/sciencenow/assets/2011/04/13/sn-language.jpg).

One of the most significant events is the migration of Niger-Congo Bantu speaking agricultural populations from the region that is known as Cameroon today initially into the rainforests of equatorial Africa, followed by separate phases of expansion into eastern and southern Africa; known as the Expansion of the Bantu Speaking Peoples (Tishkoff and Verrelli 2003; Beleza et al. 2005; Campbell and Tishkoff 2010). Genetic data also suggest that Bantu speakers migrated from east and west Africa into the southern region, see Figure 1.3. As these populations migrated they displaced many of the regional, indigenous populations and patterns of migration can be traced using mitochondrial DNA and Y chromosome markers (Quintana-Murci et al. 1999; Tishkoff and Verrelli 2003; Beleza et al. 2005; Berniell-Lee et al. 2009). The wide distributions of Niger-Congo A and Bantu languages within sub-Saharan Africa (Figure 1.2) are attributed to these migration events.

Sub-Saharan African populations also practise a wide range of subsistence methods (Campbell and Tishkoff 2008; Campbell and Tishkoff 2010). Previous studies have observed multiple mutations, at high frequencies, that have evolved to aid dietary adaptation in populations from the sub-continent (Hollox et al. 2001; Ingram et al. 2007; Perry et al. 2007; Tishkoff et al. 2007; Ingram et al. 2009; Ingram et al. 2009).

From a medical perspective, over 90% of the global burden of disease is found in developing countries (Sgaier et al. 2007; Oliveira et al. 2009) and a substantial number of developing countries are within sub-Saharan Africa. In addition to common infectious diseases such as influenza and bacterial meningitis, sub-Saharan Africa has an added burden of high rates of transmission of infectious diseases such as malaria, HIV and tuberculosis as well as so called “neglected diseases” including Leishmaniasis, Human African Trypanosomiasis and Leprosy. Approximately 800 million people reside in sub-Saharan Africa and are at risk from common and neglected diseases (Aspray et al. 1998; Hotez and Kamath 2009). Therefore it is important to understand not only how socio-economic factors may impact disease burden within the sub-continent; but also identify genetic influences that may impact disease progression, transmission and treatment.

Figure 1.3.: A map showing the phases of expansion of the Bantu Speaking People.

Image has been taken from (http://en.allexperts.com/e/b/ba/bantu.htm).

There are marked differences that have been identified between sub-Saharan African populations and European populations in the response to treatment that is administered for specific diseases (Fellay et al. 2005; Mouly et al. 2005). However over 95% of drug development and clinical trials are carried out in populations with recent European ancestry from Europe and North America. Many developing countries, including many in sub-Saharan Africa, rely on FDA and European guidelines for safety levels and optimal therapeutic dosages (Li et al. 2011). Given the substantial burden of disease within sub-Saharan Africa, coupled with high levels of genetic diversity in the sub-continent, it is necessary to examine how variation in response to treatment impacts clinical outcomes within populations from the region. This has become particularly important in recent years given the new emphasis on the use of mass drug administration (MDA) by organisations such as the World Health Organisation (WHO) and Médecins Sans Frontières (MSF) in the treatment of many endemic diseases (Hotez 2009; Smits 2009; Solomon et al. 2009). It is necessary to examine the extent of variation in the response to specific treatments and whether alternative dosages or drugs are required than those administered in different parts of the continent.

Despite these convincing reasons for including sub-Saharan Africans, they have been largely underrepresented in human evolutionary and medical studies (Campbell and Tishkoff 2008). There has been a tendency within the literature to extrapolate data obtained for African-American individuals to sub-Saharan African populations. This is problematic given the extent of European admixture within African-American populations (Reed 1969; Parra et al. 1998; Destro-Bisol et al. 1999). Additionally this method of extrapolation does not acknowledge the high level of genetic diversity between different sub-Saharan African populations, for example the high levels of Semitic admixture within East African populations (Hammer et al. 2000; Tishkoff and Verrelli 2003; Lovell et al. 2005) have not been observed in populations from western and southern Africa. Clearly there is a need for focused studies in sub-Saharan African populations to guide the effective global prescription of drugs.

1.1.5 Sub-Saharan Africa and modern human origins

The majority of archaeological and genetic data support a recent African origin model of the evolution of anatomically modern humans. This model postulates that anatomically modern humans (*Homo sapiens*) evolved in Africa from archaic humans ~150,000-200,000 years ago. Anatomically modern humans then migrated from Africa and inhabited other parts of the world replacing archaic humans with little or no genetic mixing (Quintana-Murci et al. 1999; Tishkoff and Williams 2002; Relethford 2008).

The discovery of fossil remains with early human morphological traits within sub-Saharan Africa provides support for this model of human evolution (Campbell and Tishkoff 2008; Relethford 2008). Fossil remains, with morphological features of anatomically modern humans, dated to 160,000 (White et al. 2003) and 195,000 years ago (McDougall et al. 2005) have been found in the sub-continent. These dates precede those estimated for fossils discovered in the Middle East (~92,000 years ago), Australia (~60,000-40,000 years ago) and Europe (~40,000-30,000 years ago) (Relethford 2008).

Genetic data are also consistent with a recent African origin model of human evolution (Campbell and Tishkoff 2008; Relethford 2008). Coalescent models have been used to estimate the time and location of the most recent common ancestor of modern humans. The models assume that in any sample of DNA markers, there is a point backwards in time in which they will coalesce to a common ancestor. Analyses of mitochondrial DNA, inherited solely maternally, have dated a most recent common ancestor to ~200,000 years within Africa (Relethford 2008).

High levels of genetic diversity observed within sub-Saharan Africa comparative to other global regions also supports evidence for a recent African origin of modern humans (Watkins et al. 2001; Manica et al. 2007; Campbell and Tishkoff 2008). Sub-Saharan African populations have higher levels of structural variation in the genome (Relethford 2008; Campbell and Tishkoff 2010). Additionally, sub-Saharan African populations have been reported to have more "private" alleles comparative to other global populations.

Modern humans in Africa would have a longer period of time to accumulate mutations than those who were dispersing out of the continent. The number of modern humans leaving Africa would be a small percentage of the total population; and represent a fraction of all African genetic diversity. In other words, non-African populations have been subject to a population bottleneck and so have less variation than the founding African population from which they have separated (Manica et al. 2007; Relethford 2008; DeGiorgio et al. 2009).

1.1.6 Ethiopia and human evolutionary history

East Africa is an important region in human evolutionary history. It is widely believed to be the region from which anatomically modern humans left Africa to expand across the world. Evidence for an East African migratory route, via Ethiopia, out of sub-Saharan Africa is supported by fossil evidence for anatomically modern humans dated to ~150,000-160,000 years ago (White et al. 2003; Kivisild et al. 2004).

Archaeological data support the presence of anatomically modern humans within Ethiopia (White et al. 2003; McDougall et al. 2005). Additionally, fossils supporting the presence of early hominids within the country have also been uncovered, an example is the famous "Lucy": a well preserved skeleton which was found in Ethiopia and dated to 3.2 million years ago (Shreeve 1994).

Data from studies on mitochondrial and Y chromosome DNA have successfully tracked human migration out of Africa via Ethiopia (Quintana-Murci et al. 1999; Tishkoff and Verrelli 2003; Lovell et al. 2005; Campbell and Tishkoff 2008). Additionally, global analyses of microsatellite variation (Ramachandran et al. 2005) have shown that genetic diversity is negatively correlated with distance from Ethiopia (Prugnolle et al. 2005; Li et al. 2008). In other words, the further the distance a population is geographically located from Ethiopia, the less genetic diversity it contains.

East Africa is a particularly diverse region of sub-Saharan Africa (Tishkoff et al. 2009). Genetic diversity within the region has been shaped by multiple demographic events. One of the most significant is admixture with individuals of Semitic origin who migrated into Africa via East Africa and mixed with the indigenous populations ~5000 years ago (Hammer et al. 2000; Tishkoff and Verrelli 2003; Lovell et al. 2005). This is reflected at the cultural, genetic and linguistic levels with the presence of Semitic speaking East African groups that claim Jewish ancestry, a feature that distinguishes them from other sub-Saharan African populations (Lovell et al. 2005).

1.1.7 The rationale for studying human genetic variation in Ethiopian populations

Given the high levels of genetic diversity observed in Ethiopia (Prugnolle et al. 2005; Li et al. 2008), it is anticipated that populations from this region will have higher levels of variation in medically important genes, comparative to that which has already been reported for other global populations.

A recent study on human genetic variation in the gene encoding the drug metabolising enzyme CYP1A2 in Ethiopian populations, significantly increased existing knowledge of variability in this gene (Browning et al. 2010). In addition, the authors ascertained that uncharacterised variation in medically important genes is likely to put individuals with recent Ethiopian ancestry, both within Ethiopia and overseas, at increased risk of adverse drug reactions.

Through studying variation in medically important genes, in the diverse populations of Ethiopia, and combining the results with those previously reported for other global

populations, a more complete picture of global variation in particular genes will be obtained. Additionally improved understanding of medically important variation will aid in tailoring healthcare to particular populations, and reduce the likelihood of adverse treatment outcomes; which are major contributors to annual global morbidity and mortality (Weinshilboum 2003; Wilke et al. 2007).

1.2 Part II: CYP3A5, population genetics and human adaptation

1.2.1 The Cytochrome P450 super-family of drug metabolising enzymes

Cytochromes P450 (CYP450) are a super-family of haem-containing mono-oxygenases which are widely conserved across species (Nebert and Russell 2002). There are 116 identified human *CYP450* genes: 57 encode active CYP450 enzymes; the remaining 59 are pseudogenes (http://drnelson.utmem.edu/human.P450.table.html). Human CYP450 are predominantly membrane-bound proteins, located in the inner membrane of the mitochondria or in the endoplasmic reticulum of cells (Nelson 2009). CYP450 are mainly found in the liver, although extra-hepatic isoforms exist (Nelson 2009), and are involved in the metabolism of multiple endogenous and exogenous compounds (Porter and Coon 1991). CYP450 mediate oxidation, reduction, and hydrolysis reactions which expose or add functional groups to substrates to produce polar molecules (Li et al. 2011). It is thought that the ability of CYP450 enzymes to metabolise exogenous compounds evolved 400-500 million years ago to enable animals to digest chemicals in plants, creating water-soluble compounds which are easier to excrete (Gonzalez and Gelboin 1994).

CYP450 are also important in the first phase metabolism of many therapeutic drugs used to treat a wide spectrum of diseases (Brockmoller et al. 2000; Pirmohamed and Park 2003; Kirchheiner and Seeringer 2007; Lynch and Price 2007). Of the 57 active CYP450 enzymes, six are together involved in the metabolism of more than 90% of clinically used drugs (Pirmohamed and Park 2003; Lynch and Price 2007). Not unexpectedly, polymorphisms in genes encoding CYP enzymes are associated with many adverse drug reactions (Ingelman-Sundberg 2004).

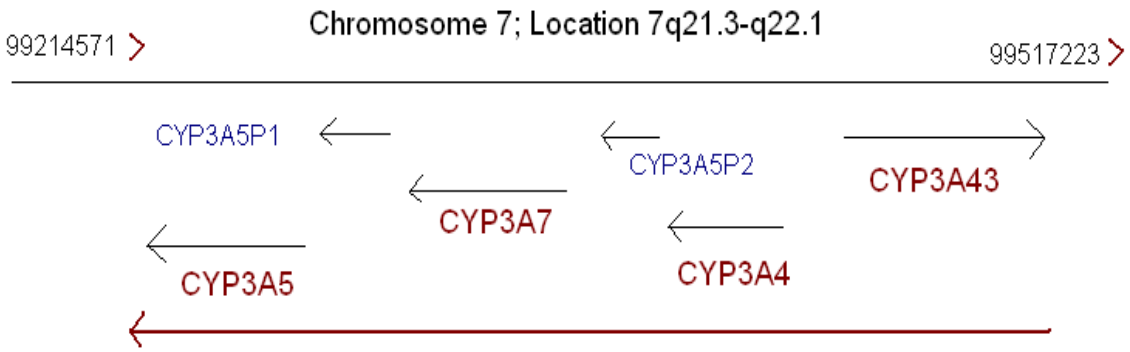
1.2.2 The Cytochrome P450 3A sub-family

The *CYP450* super-family of genes are grouped into families and sub-families; *CYP* families, such as *CYP3*, are defined by at least 40% amino acid sequence similarity; and sub-

families by 55% (Nelson et al. 2004; Qiu et al. 2008). Of the various sub-families that comprise the *CYP450* super-family, the Cytochrome P450 3A sub-family (*CYP3A*) have a central role in drug metabolism; CYP3A proteins metabolise many therapeutic drugs and are the most abundant sub-family expressed in the liver, comprising up to 50% of hepatic CYP content (Kuehl et al. 2001; Lamba et al. 2002; Xie et al. 2004). The two most important drug metabolising enzymes within this sub-family are CYP3A4 and CYP3A5 (Kuehl et al. 2001; Lamba et al. 2002; Wojnowski 2004; Xie et al. 2004). Given the wide substrate range of CYP3A, functionally significant variation in the genes encoding CYP3A enzymes can have important medical implications due to the large number of substrates metabolised solely or predominantly by this sub-family.

The *CYP3A* sub-family is mapped to a ~231kb region on chromosome 7 (Figure 1.4). There are six *CYP3A* genes; mRNA sequences corresponding to each of four *CYP3A* genes; *CYP3A4, CYP3A5, CYP3A7* and *CYP3A43* have been isolated (Lamba et al. 2002). There are also two pseudogenes: *CYP3A5P1* and *CYP3A5P2* which have sequence homology to *CYP3A5* (Lamba et al. 2002). Each of *CYP3A7* and *CYP3A4* encode proteins consisting of 503 amino acids, *CYP3A43* encodes a 504 amino acid protein and *CYP3A5* a 502 amino acid protein (http://www.ncbi.nlm.nih.gov/). There is high sequence homology at the *CYP3A* cluster, and it has been suggested that multiple *CYP3A* genes arose as a result of duplication of an ancient *CYP3A* sequence, estimated to be approximately 40-50kb in size (Finta and Zaphiropoulos 2000).

Figure 1.4: A schematic of the organisation of the human *CYP3A* locus on chromosome 7. Arrows indicate the orientation of each gene in the *CYP3A* cluster.



The *CYP3A* locus is highly conserved across mammals: *CYP3A* orthologues have been identified in primates (Williams et al. 2004; Williams et al. 2007), horses (Schmitz et al. 2010)

and in rodents (Zaphiropoulos 2003). Allelic variation and subsequent *CYP3A* haplotype diversity are high within and between species (Thompson et al. 2006; Chen et al. 2009).

For the human *CYP3A* locus, multiple *CYP3A* alleles, both functionally important and unimportant, have been identified (http://www.cypalleles.ki.se/index.htm). Different studies have reported evidence of selection on *CYP3A* genes. A paper by (Thompson et al. 2004) reported that high frequencies of a low/non-expresser *CYP3A5* variant, *CYP3A5*3,* in non-African populations were evidence of positive selection acting to increase frequencies of this derived allele outside of the African continent. However, a paper in 2006 argued that selection was acting on the functionally ambiguous *CYP3A4*1B* allele in non-African populations. As this allele has been reported to be in high linkage disequilibrium with *CYP3A5*3*, the authors argued that *CYP3A4* is a more important gene of the *CYP3A* sub-family, and was likely to be under stronger selective pressure than *CYP3A5* (Schirmer et al. 2006).

A more recent study of the *CYP3A* locus reported evidence of purifying selection on *CYP3A4* and *CYP3A7*. The authors reported low levels of nucleotide diversity, i.e. high levels of sequence conservation in the coding regions of these two genes. In contrast, the authors reported a significant departure from neutrality in the coding regions of *CYP3A5* and *CYP3A43* in Caucasian individuals, consistent with a selective sweep and positive selection. The authors reported higher frequencies of derived, non-functional *CYP3A5* and *CYP3A43* alleles in Caucasian individuals comparative to Africans (Chen et al. 2009). Recently, an examination of functionally important SNPs in multiple genes involved in drug metabolism provided further, strong evidence of a selective sweep/positive selection, on the low/non-expresser *CYP3A5*3* mutation in populations from the Middle East, Europe and Central South Asia (Li et al. 2011).

Given the substantial overlap in CYP3A drug and environmental substrates, it is difficult to ascertain exactly what may be the underlying causes of different selective patterns on individual *CYP3A* genes. Inter-population variability in haplotype structure at the *CYP3A* locus is likely to contribute to population differences in drug disposition and metabolism (Thompson et al. 2006; Chen et al. 2009).

This thesis focuses on characterising human genetic variation in the *CYP3A5* gene. Whilst many previous studies have argued that the CYP3A4 enzyme has the most significant role, of all CYP3A enzymes, in therapeutic drug metabolism (Boobis et al. 1996) reports indicate that CYP3A5 has a greater role in populations with recent African ancestry (Roy et al. 2005; Mirghani et al. 2006; Quaranta et al. 2006). Additionally the high frequencies of low/non-expresser variants in populations outside of Africa suggest that there has been differential selection on the gene between Africans and non-African populations.

1.2.3 Cytochrome P450 3A5 (CYP3A5)

Human *CYP3A5* has been mapped to the negative strand of the long arm of chromosome 7 at 7q21.1 at chromosomal location 7:99245928-99277519 (Build 132, NCBI: http://www.ncbi.nlm.nih.gov/). The gene is ~33kb long and has thirteen exons and twelve introns (Figure 1.5). One hundred nucleotides of exon 1 and the downstream sequence of exon 13 are untranslated regions. 87 nucleotides of exon 1 and 114 nucleotides downstream of exon 13 are untranslated from the mature mRNA. Multiple *CYP3A5* transcripts have been identified (NCBI: http://www.ncbi.nlm.nih.gov/), a transcript of 1509 nucleotides is translated into a protein of 502 amino acid residues (Aoyama et al. 1989). The crystallised structure of CYP3A5 has not been elucidated, unlike for its paralogue CYP3A4 (Williams et al. 2004). CYP3A5 is the major extra-hepatic CYP3A isoform; protein is also expressed in the kidneys, small intestine, prostate and lungs (Anttila et al. 1997; Lamba et al. 2002; Hukkanen et al. 2003; Moilanen et al. 2007). Additional transcripts have been identified, from cancer cell lines, although they do not encode functional nor alternative enzyme forms (NCBI: http://www.ncbi.nlm.nih.gov/).

1.2.3.1 The genetic basis of variable CYP3A5 protein expression

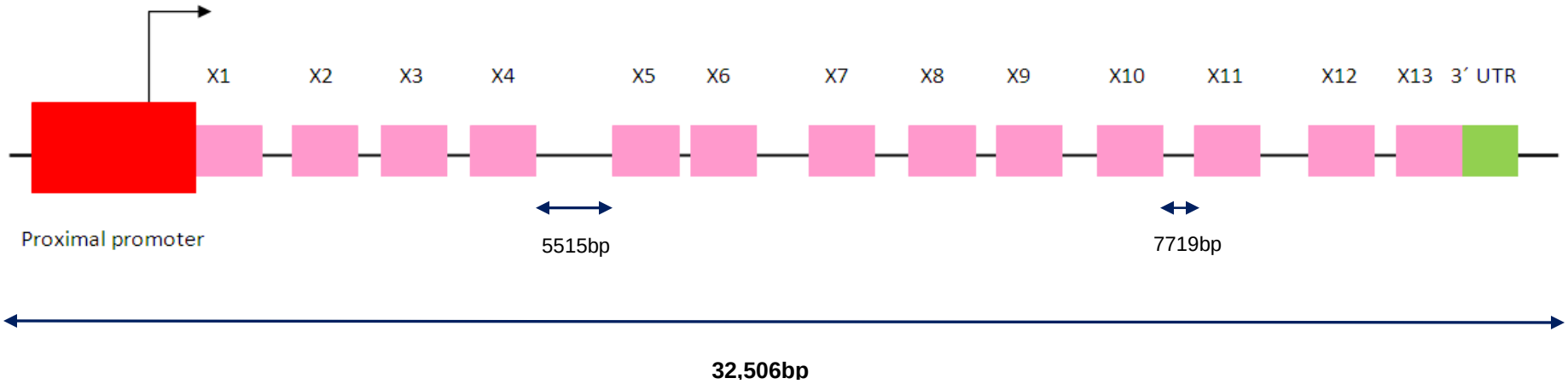
Human hepatic CYP3A5 was first detected by Aoyama and Schuetz in 1989 who observed that cDNA, corresponding to functional CYP3A5, could not be isolated from every individual and therefore enzyme expression was polymorphic; individuals tend to either express high quantities of CYP3A5 or have significantly reduced, often undetectable, protein levels (Aoyama et al. 1989; Schuetz et al. 1989).

CYP3A5 is polymorphically expressed between and within ethnic groups. Approximately 10-25% of Europeans, 30-50% of Asian and South Americans and 55-95% of African Americans (Kuehl et al. 2001; Quaranta et al. 2006) have detectable levels of hepatic CYP3A5 protein. Multiple variants have been reported to affect CYP3A5 expression (Xie et al. 2004). Four *CYP3A5* alleles; *CYP3A5*1*, *CYP3A5*3*, *CYP3A5*6* and *CYP3A5*7*, are the most common determinants of interethnic variability in protein expression (Hustert et al. 2001; Kuehl et al. 2001; Lee et al. 2003).

Figure 1.5: A representation of the *CYP3A5* locus.

Pink boxes represent exons; X*n* indicates the exon number in the sequence. The red box at the 5′ of the sequence represents the proximal promoter. The green box is the 3′ untranslated region (UTR). The arrow indicates the direction of transcription. Black lines flanking the boxes are intronic sequences. Spacing between the exons is proportional to the distance between them on the chromosome. The largest *CYP3A5* introns are annotated on the Figure.

This Figure complements Appendix A (on CD) which is the full genomic reference sequence of the *CYP3A5* gene.



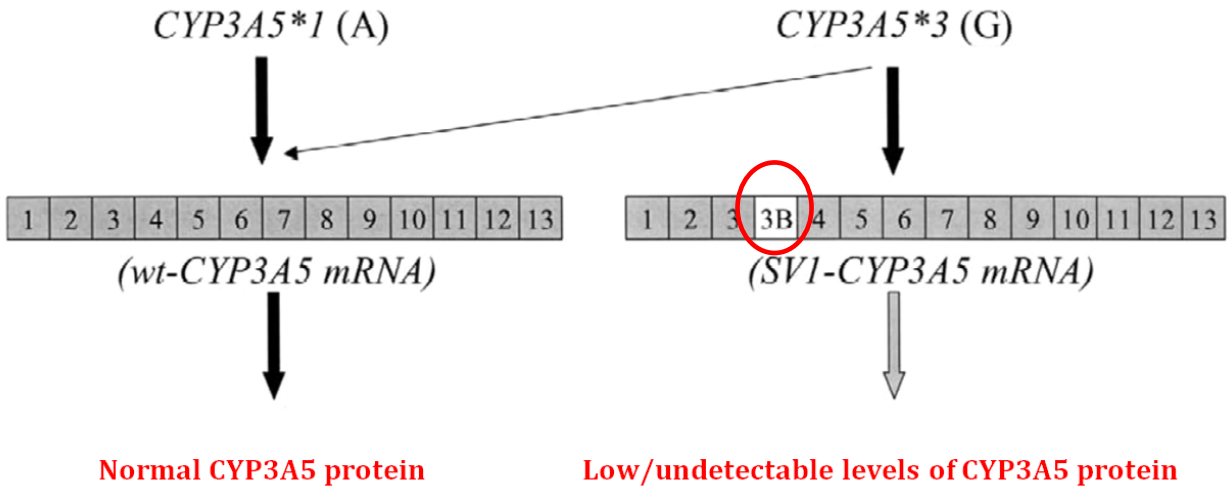
*CYP3A5*1* is the *CYP3A5* "expresser" allele and is genotyped at the same locus as *CYP3A5*3* (rs776746). Genotyping of an A allele at the *CYP3A5*1/*3* locus defines *CYP3A5*1* and a G allele defines *CYP3A5*3*. High frequencies of *CYP3A5*1* have been observed in all, genotyped, populations with recent African ancestry (Hustert et al. 2001; Kuehl et al. 2001; Wojnowski et al. 2004; Xie et al. 2004; Roy et al. 2005; Mirghani et al. 2006; Quaranta et al. 2006). Individuals homozygous for *CYP3A5*1* have been found to have levels of *CYP3A5* mRNA that make up at least 50% of total hepatic CYP3A content (Kuehl et al. 2001).

*CYP3A5*3* (rs776746; 6980A>G)[1] defines a splice variant in which an A>G transition in the third intron of *CYP3A5* causes 132 nucleotides of intron three, which would normally be spliced out, to be retained in the mature mRNA, causing a frameshift, protein truncation and a reduction in protein expression to undetectable levels, see Figure 1.6 (Kuehl et al. 2001). The original 2001 study, which identified this variant, found a significant reduction in the concentration of CYP3A5 protein in liver specimens of *CYP3A5*3* homozygotes (>21pmol/mg) than in individuals with at least one *CYP3A5*1* allele (21-202pmol/mg). *CYP3A5*3* homozygotes also had a 2.5-fold reduced ability to metabolise midazolam (a known CYP3A substrate) to its primary metabolites (Kuehl et al. 2001).

An independent study established that *CYP3A5*3* mRNA transcripts are degraded by nonsense mediated decay (Busi and Cresteil 2005). Transcripts which contain the aberrant insertion (*SV1-CYP3A5*-mRNA) are only observed in the tissues of *CYP3A5*3* carriers. *CYP3A5*3* homozygotes have high levels of *SV1-CYP3A5*-mRNA, comparative to normally spliced *CYP3A5* mRNA (Lin et al. 2002; Busi and Cresteil 2005). *SV1-CYP3A5*-mRNA levels vary between *CYP3A5*1/CYP3A5*3* heterozygotes, and this may explain inter-individual differences in CYP3A5 protein expression levels observed in *CYP3A5*3* heterozygotes (Lin et al. 2002).

[1] 6980 refers to the position, in base pairs, of the *CYP3A5*3* allele relative to the ATG start codon; consistent with all previous studies on this *CYP3A5* variant.

Figure 1.6: A diagrammatic representation of the alternative splicing pathway created by the CYP3A5*3 mutation. An A>G transition (alleles are shown in brackets) causes 132 nucleotides of intron 3 (labelled 3B and annotated on the right hand image) to be retained in the mature mRNA transcript; known as *SV1-CYP3A5-mRNA*. This causes a frameshift and premature termination of the reading frame. Normally spliced mRNA is detected in the tissues of *CYP3A5*3* homozygotes but at very low/undetectable levels. The image has been adapted from (Lamba et al. 2002).

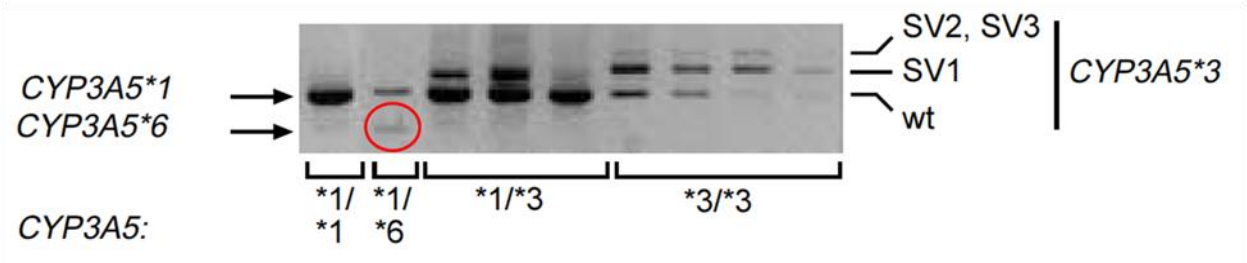


*CYP3A5*6* (rs10264272; 14684G>A) defines a G>A transition associated with exon 7 "skipping". The exonic G>A transition is a synonymous change, reported to lead to the creation of a cryptic splice site within exon 7; resulting in its excision from the mature mRNA. The exonic excision leads to a frameshift and the creation of a premature termination codon, leading to degradation of the mRNA transcript and a reduction in CYP3A5 expression levels (Kuehl et al. 2001). It was hypothesised that the *CYP3A5*6* defining G>A transition in exon 7 could disrupt an exonic splicing silencer or activate an exonic splicing enhancer, leading to an aberrant splicing pathway (Kuehl et al. 2001), although this is yet to be confirmed experimentally.

It is important to note that the effect of *CYP3A5*6* on mRNA splicing has only been established once; mRNA was extracted from three *CYP3A5*1/CYP3A5*6* heterozygotes, reverse transcribed to cDNA and the products analysed by gel electrophoresis, see Figure 1.7. For *CYP3A5*1/CYP3A5*6* heterozygotes, two cDNA products were isolated; one of which was smaller than the normal mRNA length and did not contain the sequence for exon 7 (Kuehl et al. 2001). An examination of the gel reported in the original paper (Figure 1.7) suggests that an alternative mRNA transcript is produced in *CYP3A5*6* carriers, however the results from

this paper have never been independently replicated and therefore alone are not entirely conclusive. The paucity of data on mRNA splicing and composition in *CYP3A5*6* homozygotes means that the exact effect of *CYP3A5*6* on *CYP3A5* splicing is not fully understood.

Figure 1.7: The Figure from the original paper reporting that the CYP3A5*6 allele causes aberrant splicing of CYP3A5 mRNA transcripts. mRNA from human livers was reverse transcribed and analysed by gel electrophoresis (shown below). The authors reported that the *CYP3A5*6* mRNA transcript did not have the sequence for exon 7 of the gene but did not show the sequence. The Figure has been adapted from (Kuehl et al. 2001). The circled band is the *CYP3A5*6* aberrant splice variant identified in the 2001 study. *Wt* refers to normally spliced mRNA, *SV* refers to splice variants. Data for *CYP3A5*3* heterozygotes and homozygotes are also shown on the right hand side of the image; and correspond to splice variants specifically identified for this *CYP3A5* variant.



*CYP3A5*7* (rs41303343; 27125_27126insT) defines a T nucleotide insertion into exon 11. *CYP3A5*7* is associated with a reduction in CYP3A5 expression levels. The T nucleotide insertion causes a frameshift, the creation of a premature stop codon and early termination of the open reading frame (Chou et al. 2001; Hustert et al. 2001).

Normally spliced mature mRNA, as well as abnormally spliced mRNA, is detected in all tissues in individuals heterozygous and homozygous for *CYP3A5*3* (although normally spliced mRNA levels are significantly lower, and often undetectable, in *CYP3A5*3* homozygotes) and individuals heterozygous for *CYP3A5*6*. It is important to note that unlike *CYP3A5*3* and *CYP3A5*6,* no experimental work has examined *CYP3A5* mRNA levels, or composition, in *CYP3A5*7* carriers. However, as the mutation leads to the creation of a premature stop codon, it is highly likely that these individuals do not produce any normally spliced *CYP3A5* mature mRNA unless they are *CYP3A5*1/CYP3A5*7* heterozygotes.

Many previous studies have described *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* as "knock-out", or null, *CYP3A5* alleles. However the effects of *CYP3A5*3* and *CYP3A5*6* on *in vivo* splicing have been shown to be "leaky" given the identification of normally spliced mRNA in the tissues of *CYP3A5*3* and *CYP3A5*6* carriers. (Kuehl et al. 2001; Lin et al. 2002). Many previous studies have also tended to group *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* to infer population frequencies of non-expressers of CYP3A5. However the, independent, comparative effects of

these *CYP3A5* variants on *CYP3A5* splicing and protein expression have not been experimentally established. Therefore, it is preferable to group the *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles as variants which can contribute to polymorphic CYP3A5 expression. The extensive interethnic variability in CYP3A5 expression has highlighted the need for careful sampling of individuals when evaluating the importance of this enzyme in drug metabolism (Foti and Fisher 2004; Quaranta et al. 2006).

CYP3A5 had previously been thought of as less important in the metabolism of CYP3A substrates due to its polymorphic expression and the predominance of CYP3A4 expression in the liver (Boobis et al. 1996). However, recent work has not only challenged this view but found that CYP3A5 represents at least 50% of the total hepatic and intestinal CYP3A content in individuals who express the protein at high concentrations (Lin et al. 2002). This has lead to various studies concluding that variation in the DNA sequence of *CYP3A5* may be the most important genetic contributor to interethnic and interpopulation differences in CYP3A dependent drug clearance (Kuehl et al. 2001; Lamba et al. 2002; Givens et al. 2003; Zheng et al. 2003; Foti and Fisher 2004; Frohlich et al. 2004; Xie et al. 2004; Mouly et al. 2005).

1.2.3.2 Regulation of CYP3A5 transcription

CYP3A5 transcription is regulated by a promoter situated upstream of exon 1, from -800 to +50 (Lamba et al. 2002; Burk et al. 2004). A diagrammatic representation of the proximal promoter region, including all known transcription factor binding sites, is presented in Figure 1.8. The 5´-proximal promoter region of *CYP3A5* has 89.6% homology to the equivalent region of the pseudogene *CYP3AP1* (Finta and Zaphiropoulos 2000). High levels of homology between *CYP3A5* and *CYP3AP1* lead to the initial characterisation of the *CYP3A5* promoter region being incorrect (Jounaidi et al. 1994). Given the differences in CYP3A5 expression levels between different populations, it was hypothesised that promoter variants were responsible for the variability observed in enzyme activity. A study in 2000 identified two variants, believed to be in the 5´ promoter of *CYP3A5*, as being responsible for the variable activity (Paulussen et al. 2000). However, subsequent reports found that these variants actually occurred in the 5´ upstream region of *CYP3AP1* and so could not be responsible for polymorphic CYP3A5 expression (Lamba et al. 2002). A later study did find that linkage disequilibrium (LD) between these variants and the *CYP3A5*1* allele is high (Kuehl et al. 2001) although this is likely to be due to high levels of LD across the entire *CYP3A* cluster (see chapter 5) (Thompson et al. 2006).

To date, few variants have been identified in the *CYP3A5* proximal promoter (Nakamura et al. 2001); none occur in known transcription factor binding sites of the gene;

and none of the variants have been reported to be associated with low or altered CYP3A5 enzyme activity (Lamba et al. 2002; Xie et al. 2004). Due to high levels of LD across the gene, these variants are often found on the same haplotype background as the *CYP3A5*3* allele (Xie et al. 2004).

CYP3A5 is not as inducible by nuclear hormones/genes as *CYP3A4* (Lamba et al. 2002; Lin et al. 2002) and transcription is not known to be regulated by a distant enhancer located ~7000 base pairs upstream of the *CYP3A* cluster, as is the case for *CYP3A4* (Matsumura et al. 2004; Martinez-Jimenez et al. 2005). However, CYP3A5 is the major extra-hepatic CYP3A enzyme (Lamba et al. 2002; Wojnowski 2004) and a previous study reported that CYP3A5 accounts for at least 50% of the CYP3A hepatic content in individuals who are homozygous for *CYP3A5*1* (Kuehl et al. 2001). *CYP3A5* transcription can be induced by a number of transcription factors including specificity proteins (Sp) and nuclear factor-Y proteins (Iwano et al. 2001); both classes of these transcription factors are ubiquitous in human tissues (Roder et al. 1999; Suske 1999; Kolell and Crawford 2002) and this may explain the large tissue range in which the *CYP3A5* gene is transcribed and CYP3A5 enzyme is expressed.

1.2.3.3 The 3´ downstream region of CYP3A5

The 3´ region of a gene is a critical regulator of protein expression. The 3´ region encodes a series of signals which can affect translation termination, the export from the nucleus and stability of an mRNA transcript (Neilson and Sandberg 2010). Polymorphisms which occur in the 3´ untranslated region can significantly affect protein expression and cellular localisation (Alt et al. 1980). Additionally, a number of mutations which occur in the 3´ region of specific genes have been associated with several disease pathologies, including numerous cancers (Chatterjee and Pal 2009). Several *CYP3A5* mRNA transcripts, with alternate 3´ ends, have been isolated from cancer cell lines (http://www.ncbi.nlm.nih.gov/). Despite the importance of the 3´ region, studies have estimated that over half of all human genes have multiple mRNA 3´ ends (Tian et al. 2005; Hughes 2006) suggesting that eukaryotic gene regulation is much more complex than by the 5´ region alone.

The 3´ region of the full *CYP3A5* transcript, known to encode a 502 amino acid protein, is 114 base pairs in length. To date three variants have been identified in the 3´ UTR of *CYP3A5* (http://www.ncbi.nlm.nih.gov/) and of these two occur at global frequencies of over 1%. One of these variants, rs15524, is tightly linked to the *CYP3A5*3* allele; although experimental work has found that this variant alone does not influence *CYP3A5* mRNA processing, nor does it impact protein expression in any other way (Busi and Cresteil 2005).

Figure 1.8: The known proximal promoter region of *CYP3A5*. Known regulatory motifs are in red text and underlined. The diagram has been adapted from (Lin et al. 2002). Positions are numbered according to the ATG start codon where base A is +1.

Abbreviations: NFSE, Nifedipine specific element; PRE/GRE, pregnane/glucocorticoid receptor element; ERE, oestrogen response element; HNF-5, human nuclear factor-5; PXR-RE (ER-6), pregnane X receptor element with a six nucleotide everted repeat; BTE, basal transcription element.

1.2.4 Drugs metabolised by CYP3A5

Variability in CYP3A5 expression has a significant effect on drug treatment-related adverse clinical outcomes and disease susceptibility. A significant proportion of CYP3A substrates are metabolised by both CYP3A4 and CYP3A5. These enzymes overlap in their substrate specificity, however studies have shown that overlapping substrates are metabolised much more efficiently in individuals who express both CYP3A5 and CYP3A4 than in individuals who solely express CYP3A4 (Lin et al. 2002; Zheng et al. 2003; Mouly et al. 2005; Mirghani et al. 2006). It should be noted that it is unclear whether an increase in metabolism is due to the specificity of CYP3A5 for certain CYP3A substrates, or due to the presence of two CYP3A enzymes that are equally efficient at metabolising CYP3A substrates and hence enable much more efficient metabolism. However studies have found a significant contribution of CYP3A5 in the metabolism of HIV-1 protease inhibitors, drugs used in the treatment of severe malarial infection, kidney diseases and mental illness. In each of these examples, the contribution of CYP3A5 to drug metabolism was greater than CYP3A4 (Kuehl et al. 2001; Lin et al. 2002; Givens et al. 2003; Zheng et al. 2003; Frohlich et al. 2004; Wojnowski et al. 2004; Mouly et al. 2005; Mirghani et al. 2006; Kohlrausch et al. 2008).

1.2.4.1 HIV-1 protease inhibitors

One of the most important classes of drugs that are metabolised by CYP3A5 are protease inhibitors used in the treatment of HIV-1 infections (Haas et al. 2006; Josephson et al. 2007); one of the best studied examples in pharmacogenetics is saquinavir (Frohlich et al. 2004; Mouly et al. 2005; Josephson et al. 2007).

Saquinavir is a HIV-1 protease inhibitor and has been reported as having low bioavailability (the proportion of a drug that is available to the target body tissue after administration) due to extensive first phase metabolism by Cytochromes P450 (Frohlich et al. 2004; Mouly et al. 2005). Clinical studies have also found that saquinavir bioavailability has interethnic and interpopulation variability; a feature that is consistent with CYP3A substrates.

Recent studies have found that saquinavir bioavailability is influenced by hepatic CYP3A5 content (Frohlich et al. 2004; Mouly et al. 2005; Josephson et al. 2007). In one study, urine samples were tested for parent/metabolite ratio. Increased bioavailability would result in a higher proportion of metabolite than parent compound. The study found that the ratio was independent of CYP3A4 activity, however in individuals genotyped with at least one *CYP3A5*1* allele; hepatic expression of CYP3A5 correlated positively with a higher proportion of metabolite in the urine samples and furthermore the proportion was twice as high as that for individuals heterozygous or homozygous for *CYP3A5*3*, i.e. low/non-expressers (Mouly et

al. 2005). Significantly, intestinal CYP3A5 content was not correlated with increased ratio of metabolite to parent compound; indicating a significant role for hepatic CYP3A5 in saquinavir metabolism.

An additional study found that patients who were homozygous for two low/non-expresser *CYP3A5*3* alleles had a 34% increased likelihood of treatment failure; individuals who were heterozygous for *CYP3A5*1/CYP3A5*3* had increased time to clear the drug from the system compared to *CYP3A5*1* homozygotes. Thus patients who were *CYP3A5*3* homozygotes were exposed to the toxic compound saquinavir for a longer period of time than patients who were CYP3A5 expressers; and so at increased risk of adverse clinical effects (Josephson et al. 2007).

Aside from saquinavir, researchers are now assessing the role of CYP3A5 in the metabolism of other drugs used in the control of HIV-1 infections, and for new anti-retroviral agents such as Maraviroc which targets the chemokine receptor CCR5 on the host cell with the aim of blocking viral entry. A recent study found that CYP3A5 appeared to have a role in the bioavailability of this drug but concluded that further work was required to determine the exact nature of the interaction (MacArthur and Novak 2008).

These studies highlight a significant role for CYP3A5 in the metabolism of drugs used in the control of HIV-1 infections. They have also provided evidence for the importance of hepatic CYP3A5 in the metabolism of CYP3A, and CYP450, specific substrates.

1.2.4.2 Drugs used in the treatment of severe malarial infections

CYP3A are involved in the first phase metabolism of drugs used in the treatment of malaria infections (Mirghani et al. 2006; Diczfalusy et al. 2008; Ferreira et al. 2008). In malaria endemic regions, the recommended first line of treatment for uncomplicated malaria is artemisinin-combination therapy (ACT) in which an artemisinin containing drug (e.g. artemether) is paired with another anti-malarial agent (Cook et al. 2009). Artemisinin has been shown to be a substrate for CYP3A4 *in vitro* and *in vivo* (Ferreira et al. 2008; Piedade and Gil 2011); few studies have examined the ability of CYP3A5 to metabolise this compound. However studies have reported a role for CYP3A5 in the metabolism of quinine; an agent used in the treatment of severe malarial infections (Mirghani et al. 2006; Diczfalusy et al. 2008; Ferreira et al. 2008).

In 2006 (Mirghani et al. 2006) explored the relationship between *CYP3A5* variation and quinine metabolism in Tanzanian and Swedish individuals. Healthy participants were given one 250mg oral dose of quinine hydrochloride and a blood sample was taken from each

participant sixteen hours post drug administration. The investigators measured the metabolic ratio of quinine hydrochloride (the parent compound) to 3-hydroxyquinine (the primary metabolite with anti-malarial properties). Individuals who were found to have high levels of 3-hydroxyquinine in their blood compared to quinine hydrochloride (i.e. had a low metabolic ratio) were considered to be efficient metabolisers of the oral drug. Each participant was genotyped for the *CYP3A5*1/CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles and inferred CYP3A5 expression phenotypes were compared to the metabolic ratio of parent compound to metabolite to examine the association.

Tanzanians had a significantly lower mean quinine metabolic ratio when compared to Swedes i.e. Tanzanians on the whole metabolised quinine hydrochloride much more effectively than Swedish participants were able to. The study also found that Tanzanians were more likely to express CYP3A5 at detectable levels, due to higher frequencies of the *CYP3A5*1* allele than observed in the Swedish group; consistent with independent reports (Hustert et al. 2001; Kuehl et al. 2001; Roy et al. 2005; Quaranta et al. 2006). Inferred low/non-expressers of CYP3A5 from both populations had high quinine metabolic ratios, i.e. low concentrations of the primary metabolite 3-hydroxyquinine but high concentrations of the parent compound quinine hydrochloride in the blood, indicating that the bioavailability of the active compound was low. Conversely, individuals who were CYP3A5 expressers had low quinine metabolic ratios, i.e. high concentrations of the primary metabolite 3-hydroxyquinine and low concentrations of the parent compound quinine hydrochloride in the blood. This suggests a role for CYP3A5 in the metabolism of this drug substrate. Interestingly the study also found that Tanzanians who did not express CYP3A5 (i.e. only CYP3A4 was expressed) were significantly less able to metabolise quinine than Swedes who only expressed CYP3A4. Since malaria is endemic in Tanzania and not Sweden, this is an important example of how population-specific expression of CYP3A5 could have significant clinical implications.

1.2.4.3 Drugs used in the treatment of mental illness

Previous research has identified a role for CYP3A enzymes in the metabolism of various drugs used in the treatment of depression and other mental illnesses (Lin et al. 2002; Floyd et al. 2003; Eap et al. 2004; Wong et al. 2004; Yu et al. 2004; Fromm et al. 2007; Kang et al. 2009). A recent paper found a significant role for CYP3A5 in the metabolism of certain drugs used to manage schizophrenia in European-Brazilian patients (Kohlrausch et al. 2008).

This study examined the influence of SNPs in five genes on the ability to respond to treatment with haloperidol and chlorpromazine in 186 European-Brazilian patients with

schizophrenia; both drugs need to be metabolised for bioavailability of active compounds. Patients were identified as non-responders to treatment if they were unable to metabolise either or both of these drugs and if they did not demonstrate appropriate behavioural control within two years of starting therapy.

A total of 40 known SNPs in five genes (two dopamine receptors; *DRD2* and *DRD3* and three Cytochrome P450 genes; *CYP2D6, CYP3A4* and *CYP3A5*) were genotyped based on evidence of their effect on protein function; a single SNP was genotyped in *DRD2,* five in *DRD3*, 24 polymorphisms in *CYP2D6*, nine in *CYP3A4* and one in *CYP3A5*.

Multiple logistic regression analysis found that individuals who had at least one copy of a haplotype of mutations at all five loci of the dopamine receptor *DRD3* and who had one copy of the *CYP3A5*3* allele had a significant reduction in the ability to respond to treatment. Individuals homozygous for the *CYP3A5*3* allele but did not have mutations in the dopamine receptor *DRD3* had a resistance to treatment, although resistance to treatment was not as severe in these patients as in those who had mutations in both of these genes.

This study was one of many that have examined the effect of variation in CYP450 on the ability to metabolise drugs used in the treatment of mental illness (Lin et al. 2002; Floyd et al. 2003; Eap et al. 2004; Wong et al. 2004; Yu et al. 2004; Fromm et al. 2007; Kang et al. 2009). However it should be noted that it is unclear as to whether *CYP3A5* variation affects the ability of an individual to absorb these drugs, into the site of action, or on the ability to excrete drugs effectively from the system. Further research will need to establish the molecular interactions between the enzyme and substrate in order to determine the exact role of CYP3A5 in the metabolism of these drugs.

1.2.4.4 Drugs used in the management of patients post solid organ transplantation

One of the best studied examples of CYP3A substrates are therapeutic drugs used in the management of patients following solid organ transplantation. When individuals undergo solid organ transplantation there is a high chance that the transplanted organ will be rejected by the patient's immune system. Following surgery it is necessary to prescribe immunosuppressive drugs to manage this immune response. Aside from problems that exist due to compliance of patients with taking these medications, there is a high degree of variability in treatment outcome following surgery; of which a significant proportion can be attributed to variation in response to immunosuppressive therapy.

One of the most commonly used drugs following solid organ transplantation is tacrolimus (Iwasaki 2007; Cattaneo et al. 2008). Tacrolimus can be administered both orally and intravenously. High rates of interethnic variability in drug treatment have been reported

for orally administered tacrolimus; a feature that is consistent with CYP3A substrates (Iwasaki 2007). Intravenously administered tacrolimus displays minimal interethnic variability and improves clinical outcome (personal communication with Dr Ian MacPhee)[2]. However intravenous administration of the drug is a lengthy and difficult process due to the need for repeated injections every four hours within hospital and after discharge of the patient; it is for this reason that oral administration of this compound is favoured over intravenous administration and studies have attempted to adjust oral concentrations of tacrolimus to improve clinical outcome. Tacrolimus is a substrate for multiple drug metabolising enzymes, including those outside of the CYP450 superfamily, however there is a significant association between the effective metabolism of the drug and CYP3A5 expression (Hesselink et al. 2003; Zheng et al. 2003; Goto et al. 2004; Zhao et al. 2005; Quteineh et al. 2008).

Tacrolimus needs to be metabolised in order to become active and enable immunosuppression. Tacrolimus has a narrow therapeutic index; i.e. there is a narrow window between concentrations of the drug that are effective and concentrations that are toxic to patients. Therefore the interethnic variability in drug response is problematic as standardised dosing of this drug following surgery has biased against groups of individuals who are much more likely to express CYP3A5 than groups who do not. Individuals who express CYP3A5 tend to metabolise tacrolimus at a much faster rate than low/non expressers, therefore they have an insufficient dosage to ensure immunosuppression and higher rates of organ rejection. Conversely, low/non expressers of CYP3A5 tend to metabolise tacrolimus at slower rates than CYP3A5 expressers which is problematic as they are exposed to a toxic compound for longer periods of time and have increased levels of liver toxicity (Hesselink et al. 2003; Zheng et al. 2003; Goto et al. 2004; Zhao et al. 2005; Iwasaki 2007; Cattaneo et al. 2008; Quteineh et al. 2008).

Therefore it has become necessary to examine how best to tailor tacrolimus treatment to an individual patient based on their *CYP3A5* genotypic profile. However this will require further studies on other factors that can affect tacrolimus metabolism as it is not understood how multiple proteins involved in tacrolimus metabolism may interact at the molecular level and whether this influences tacrolimus bioavailability (Cattaneo et al. 2008).

[2] Dr Ian MacPhee is an academic consultant nephrologist at St George's Hospital in London with a specific interest in kidney transplantation. He has published numerous papers on the pharmacogenetics of tacrolimus.

1.2.5 CYP3A5 and disease risk

Variability in CYP3A5 expression is also attributed to inter-ethnic differences in the risk of developing certain diseases, including hypertension (Fromm et al. 2005) and kidney diseases (Givens et al. 2003; Zheng et al. 2003; Quaranta et al. 2006). Controversial and contested reports have identified associations between *CYP3A5* alleles and increased risk for the development of different cancers (Dandara et al. 2005; Zhenhua et al. 2005). Elevated disease risks may be associated with the role of CYP3A5 in the metabolism of endogenous substrates including oestrogens, testosterone, androgens and steroid hormones (Lamba et al. 2002). Of all identified disease risks, the best studied example is the association between salt-sensitive hypertension and *CYP3A5* genotypes.

1.2.5.1 CYP3A5 variability and hypertension risk

Hypertension is a leading cause of global morbidity, affecting approximately 37-55% of the adult population in Europe (Wolf-Maier et al. 2003) and an even higher proportion of adults in Asia and Africa (Brown 2006). Hypertension is diagnosed based on readings of the pressure at which blood is pumped from (systolic blood pressure) and to the heart (diastolic blood pressure). Hypertension is categorised into three major categories: pre-hypertension, stage one hypertension and stage two hypertension (National Institute for Clinical Excellence [NICE], United Kingdom: http://www.nice.org.uk/newsroom/pressreleases/NewGuidelineForDiagnosingAndTreatingHighBloodPressure.jsp). A diagnosis of hypertension is made when readings of systolic and diastolic blood pressure, for an individual patient, are consistently in the range for classifying stage one hypertension.

A number of dietary and environmental factors are associated with hypertension, however more recently genetic associations have also been identified (Bochud et al. 2009). Associations between high CYP3A5 protein expression levels and hypertension have been reported (Givens et al. 2003; Fromm et al. 2005). CYP3A5 is involved in the metabolism of the endogenous substrate cortisol to 6β-hydroxycortisol, a key regulator of renal sodium transport (Wrighton et al. 1990). Selective inhibitors of CYP3A5 have been found to decrease the level of 6β-hydroxycortisol and decrease blood pressure (Watlington et al. 1992). It has been proposed that *CYP3A5*1* carriers have enhanced sodium reabsorption and, given the high frequencies of *CYP3A5*1* in populations with recent African ancestry, it has been proposed that high frequencies of this *CYP3A5* allele are advantageous in equatorial populations in times of water shortage (Kuehl et al. 2001; Thompson et al. 2004). However,

there are conflicting data on whether *CYP3A5*1* is associated with hypertension, and where associations have been observed, there are conflicting data on whether the association is to increase or decrease blood pressure (Bochud et al. 2009). The associations between *CYP3A5*1* and hypertension appear to be more pronounced in populations with recent African ancestry (Thompson et al. 2004). However there may be a possibility of ascertainment bias in studies which only sample African or African-American populations as they are known to have high frequencies of *CYP3A5*1*. Comparisons of populations with recent African ancestry and other global populations will provide a better indication of how important *CYP3A5*1* is in hypertension physiology.

1.3 Part III: Population genetic theory and selection

Population genetics examines variation in haplotype and allele frequencies to determine which evolutionary processes have occurred, or are occurring within and between populations. Mutation is the main source of genetic variation. The three main processes which can influence allele and haplotype frequencies are gene flow, drift and selection (Hedrick 2007). Population genetics models need to account for, and differentiate between, each of these processes to explain observed patterns of diversity. Of particular relevance to this thesis are methods used to detect selection.

1.3.1 The different trajectories of neutral and selected mutations

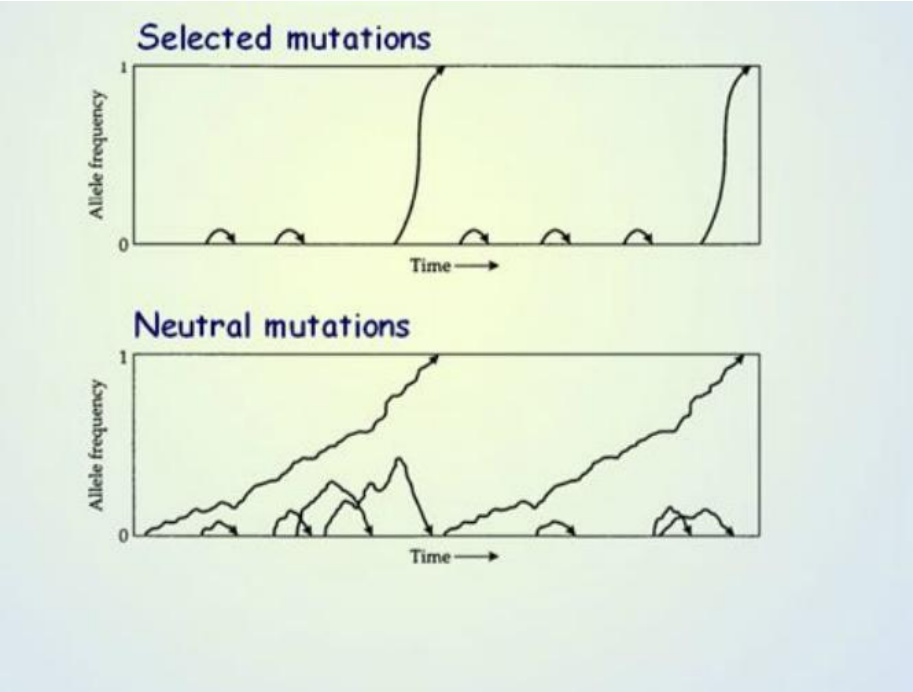
Selection may influence allele and haplotype frequencies, we can compare the trajectories of neutral and selected mutations (Figure 1.9a). The two graphs compare changes in the frequencies of selected and neutral mutations over time. The frequency of a neutral mutation is entirely dependent upon genetic drift. A neutral mutation can rise to high frequencies/fixation, but this process is much slower than for a selected allele; as seen in the top graph.

In contrast, the time taken for an allele that confers a selective advantage to increase in frequency is much lower than for neutral alleles. In either graph, new alleles are being continuously introduced into the population by mutation, but alterations in their frequencies are entirely dependent upon genetic drift; consequently many mutations remain at low frequencies or are lost from populations.

1.3.2 Signatures of positive selection within the human genome

The majority of new mutations are lost as a result of random genetic drift. The neutral theory of evolution proposes that the majority of inter- and intra-population differences in allele frequencies are a result of random fluctuations in the frequencies of neutral mutations (Kimura 1979; Kimura 1991). Population genetics models have been developed to identify marks, or "signatures", of selection within the human genome (Sabeti et al. 2007). An overview of models for detecting positive selection, which are relevant to work presented in this thesis, is presented below.

Figure 1.9a: The different trajectories of neutral and selected mutations. Image has been taken from Di Rienzo, A. (2007), "The Signature of Local Adaptations in Human Polymorphism Data": a talk from the Henry Stewart Biomedical and Life Sciences collection (www.hstalk.com/bio).



1.3.2.1 Comparing non-synonymous and synonymous mutations

Population genetics has focused on examining how variation within gene coding regions may be a result of selection. Exonic variants tend to fall into two categories: those which are predicted to alter protein structure/function (often non-synonymous/harmful mutations) and benign (synonymous) mutations.

Deleterious mutations are unlikely to reach high frequencies within a population; therefore if a non-synonymous mutation is neutral it will increase in frequency at the same rate as synonymous mutations. This can be measured by comparing the ratio of non-

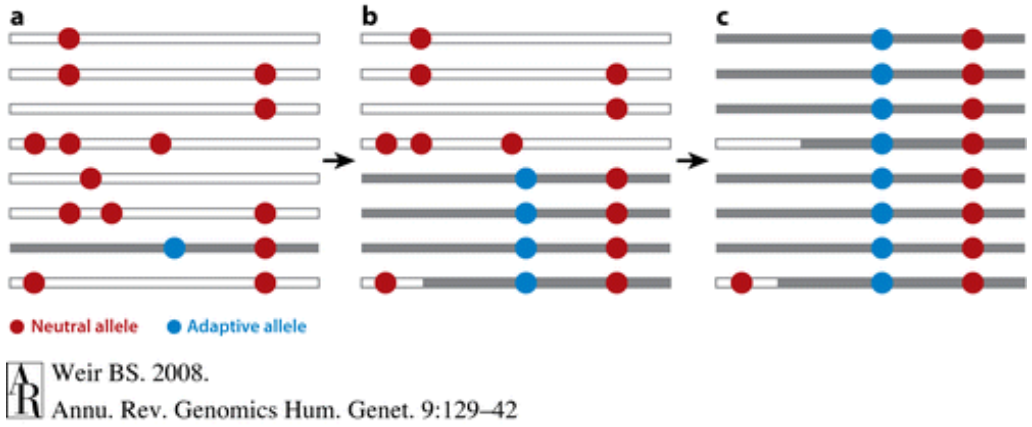
synonymous (d_N) to synonymous (d_S) changes. If a non-synonymous change is neutral then the ratio d_N/d_S will be equal to 1; if the change is deleterious then it will not reach high frequencies/fixation and d_N/d_S will be equal to less than 1; if a non-synonymous change is advantageous then d_N/d_S will be equal to greater than 1. If an excess of high frequency function-altering variants are observed in the data, which can be tested using a McDonald-Kreitman test (McDonald and Kreitman 1991), then it may indicate an adaptive change/positive selection for a particular mutation.

Whilst this method has been useful in identifying changes which can affect the expression and function of particular genes; it overlooks the role of non-coding and regulatory genomic regions which can also affect gene function and expression. An additional limitation of this test is that it relies heavily on the assumption that non-synonymous variation is function altering when this is not universally the case. However this test is useful for evaluating coding region selection.

1.3.2.2 Allele frequency spectrums

As seen in Figure 1.9a, an adaptive allele will rise to high frequency along with additional neutral variation to which it is tightly linked; this is known as a selective sweep, see Figure 1.9b. Here an adaptive allele arises in the population, is selected for and therefore increases in frequency; eventually defining all haplotypes within a population (Figure 1.9b-c).

Figure 1.9b: The mechanism of a selective sweep whereby an adaptive allele (shown in blue) increases in frequency along with all tightly linked neutral variation (shown in red), and eventually goes to fixation, as a result of a selective sweep. The image has been taken from (Kelley and Swanson 2008).



A reduction in polymorphism levels is often seen following a selective sweep (until new mutations or recombination events increase variation in and around the selected region). The overall mutation rate in the human genome is estimated to be 10^{-8} (Jobling et al. 2004) for

each nucleotide per generation. Given the rapid increase in frequency of an adaptive allele, along with all tightly linked variation, new mutations often do not accumulate on these haplotypes and those that do are rare and so observed at low frequencies (Figure 1.9). Therefore an excess of rare alleles, which can be measured by the Tajima's *D* statistic (Tajima 1989), and a reduction in polymorphism levels, which can be measured by the HKA test (Wright and Charlesworth 2004), in a genomic region may indicate a past selective event.

1.3.2.3 High frequency derived alleles

Mutation can introduce new alleles into a population. Neutral mutations can rise to high frequency as a result of genetic drift (Figure 1.8); but this process takes much longer than for selected alleles. During a selective sweep, derived alleles which are tightly linked to an adaptive allele can rise, or effectively "hitch-hike" to high frequency. Some neutral variants may not increase in frequency as a result of an incomplete sweep, or recombination of the selected region during a selective sweep. Therefore a region containing many derived alleles at high frequencies can indicate a past selective event. A significant number of high frequency derived alleles can be measured by Fay and Wu's H test (Fay and Wu 2000).

1.3.2.4 Population differentiation

Population specific differences in allele frequencies may indicate differential positive selection in one population over another. One of the most common methods for identifying population differentiation is the F_{ST} statistic (Wright 1950). F_{ST} measures the significance of population differences in allele frequencies; the statistic ranges from 0 to 1; where 1 indicates that inter-population allele frequencies are different.

However there are problems associated with forming conclusions about differential selection based on use of the F_{ST} statistic alone. Large differences in allele frequencies are estimated to occur at over 30% of all polymorphic loci identified in the human genome. Furthermore, demographic factors; in particular population bottlenecks where substantial genetic diversity is lost (Manica et al. 2007), can mimic a signature of positive selection and lead to significant inter-population differences in allele frequencies and in F_{ST} values (Hofer et al. 2009). Additionally the F_{ST} statistic will identify inter-population differences for both neutral and adaptive alleles. Therefore the test should be used in conjunction with those which examine functional implications of particular loci in order to identify adaptive alleles. Inferences about positive selection in the human genome should also use additional tests, in conjunction with the F_{ST} statistic.

1.3.2.5 Haplotype based methods for detecting selection

In addition to allele frequency tests of selection; there are also those which examine haplotype diversity and structure. As mentioned previously, a single allele can be a target of selection but it will often rise to high frequency along with a tightly linked genomic region; known as a haplotype (or large linkage disequilibrium block). Selective sweeps can be both hard and soft; and each with different signatures on haplotypes surrounding the adaptive allele.

A long region of extended haplotype homozygosity (EHH) surrounding the adaptive allele is often seen after a hard sweep. This is seen as additional mutations would not have had time to accumulate on a region tightly linked to an adaptive allele, and recombination will not have been able to rearrange genomic regions to increase diversity (Kim and Nielsen 2004). Many scans for signatures of hard selective sweeps have identified multiple regions believed to have undergone positive selection (Sabeti et al. 2002; Sabeti et al. 2006; Sabeti et al. 2007) and through the use of a long range haplotype (LRH) test identified regions of low diversity tightly linked to adaptive mutations (Sabeti et al. 2002).

However there are instances where a previously neutral allele may become advantageous, perhaps due to new environmental conditions, and so becomes adaptive and selected for. This allele may increase in frequency to fixation rapidly due to a selective sweep on standing neutral variation. Given the previous neutrality of the now selected allele, there is likely to be substantial diversity in the haplotype classes which have now risen to high frequency. This is due to recombination and mutation events during the neutral phase of the allele; additionally tightly linked regions may be smaller and so unsuitable for examination with the LRH test. A review of selective sweeps (Pritchard et al. 2010) proposed that soft selective sweeps would enable necessary adaptation and were perhaps much more prominent in our evolutionary history than hard selective sweeps alone.

1.3.3 CYP3A5 and the salt-retention hypothesis

Some previously published studies have attempted to explain the significant inter-population differences observed in the frequencies of low/non-expresser *CYP3A5* alleles. A paper by (Thompson et al. 2004) found a strong positive correlation between high frequencies of a low/non-expresser *CYP3A5* allele; *CYP3A5*3* and distance from the equator.

The authors speculated that the *CYP3A5* expresser allele; *CYP3A5*1*, may be advantageous in populations close to the equator due to the role of CYP3A5 in sodium reabsorption and water retention. CYP3A5 is responsible for the conversion of cortisol into 6β-hydroxycortisol in the kidney. 6β-hydroxycortisol is important in maintaining immune

responses that cause inflammation, and is a key regulator of renal sodium transport (Wrighton et al. 1990). A side effect of 6β-hydroxycortisol is increased sodium and water retention which is responsible for the clinical phenotype of salt-sensitive hypertension. For equatorial human populations, retention of salt and water was proposed to be advantageous; particularly during periods of water shortage. However the retention of salt in populations residing in non-humid climates is disadvantageous and so it was proposed that the correlation between high frequencies of *CYP3A5*3* and increased latitude was due to a selective advantage in non-equatorial populations to not readily retain salt as much as populations from humid climates.

In fact a strong positive correlation is observed between increased latitude and functionally important variants of genes implicated in salt-sensitive hypertension (Young et al. 2005). Given the substantial diversity observed in sub-Saharan Africa, it will be interesting to see whether there is any evidence of longitudinal differences in low/non-expresser *CYP3A5* allele frequencies between populations from the sub-continent.

1.4 Aims and overview of thesis

The overall objective of this thesis is to examine whether there are significant inter-population differences in, and evidence of positive selection on, the gene encoding the drug metabolising enzyme Cytochrome P450 3A5 (CYP3A5) in sub-Saharan Africa; with a specific emphasis on Ethiopian populations.

The specific aims addressed within each of the results chapters are as follows:

1. To determine and compare the frequencies of the three main low/non-expresser *CYP3A5* alleles: *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* and the expresser allele: *CYP3A5*1* in 36 geographically and ethnically distinct populations from in and around sub-Saharan Africa.

2. To examine intra-African diversity within a ~4.5kb region of *CYP3A5.* To compare sub-Saharan African diversity in the *CYP3A5* gene with data previously reported for global populations.

3. To identify novel variants, from sub-Saharan African re-sequencing data, that may affect *CYP3A5* transcription, translation or CYP3A5 expression.

4. To compare diversity within a ~12.3kb region of *CYP3A5* in five ethnically distinct Ethiopian populations. To then compare diversity within an overlapping re-sequencing data for an 8063 base pair region of *CYP3A5*, between Ethiopians and other global populations.

5. To characterise the evolutionary relationships between expresser and low/non-expresser *CYP3A5* haplotypes and to date low/non-expresser variants.

2. Materials and Methods

2.1 DNA samples and population histories

All samples re-sequenced and genotyped for this thesis are part of a large DNA collection at University College London, UK. The samples were collected anonymously and with informed consent (verbal in Africa) from specified locations in and around Africa [ethical approval: UCLH 99/0196]. Additional ethical approval was obtained from the Ethiopian Science and Technology Commission in Addis Ababa for Ethiopian collections. The samples were collected from males aged 18 years and over, and unrelated at the parental and paternal grandparental level. Sociological data including age, current residence, birthplace, primary language spoken, self-declared ethnic identity and religion of the individual, the individual's father and paternal grandfather were also collected. Individuals were grouped, either by the location from which they were collected or by ethnicity, into "sample sets" for analysis. The criterion for inclusion of individual samples was that they could be grouped with at least 19 other individual samples by either ethnicity and/or a shared language family within a particular location. Individuals who could not be grouped by ethnicity and/or shared language family were excluded from the studies. Samples were not grouped according to country; the partitioning of the African continent by colonial powers was recent and largely irrespective of ethnic identities (Pakenham 1991). Analysis of sub-Saharan African diversity by ethnicity, language or specified location within Africa is an appropriate method.

2.1.1 DNA samples used for re-sequencing of CYP3A5

Thirteen sub-Saharan African sample sets were chosen for re-sequencing of *CYP3A5*. The entire *CYP3A5* coding region, exon-flanking introns, proximal promoter and 5′ and 3′ gene flanking sequence, were re-sequenced in five Ethiopian populations.

2.1.1.1 Ethiopian samples

The five populations re-sequenced for this thesis overlap completely with those used to characterise genetic variation at the *CYP1A2* locus (Browning et al. 2010). The five Ethiopian groups represent a rough northeast to southwest transect across Ethiopia (Figure 4.7). Pairwise F_{ST} estimates, based on Ethiopian Y chromosome and mitochondrial DNA hypervariable region 1 data (unpublished), suggest that the majority of Ethiopian genetic variation is captured by sampling a northeast to southwest transect across the country.

2.1.1.1.1 Afar

The Afar are a pastoralist population located primarily in the Afar region, located in the northeast, of Ethiopia (Lewis 1994). The 2008 census reported that the total population size of the Afar as 1,276,372 (http://www.ethnologue.com/web.asp). The Afar language is a member of the Cushitic branch of the Afro-Asiatic language family (Lewis 1994). Given the harsh environmental conditions of the Afar region the population are nomadic pastoralists, who raise goats, sheep, camels and cattle in the desert (Getachew 1998; Getachew 2002). The pastoral and nomadic practises of the Afar are similar to Arabs (Murdock 1959) and it is likely that genetic diversity within the Afar has also been shaped by Arab migrations in and out of the region (Henze 2000; Kivisild et al. 2004; Lovell et al. 2005).

2.1.1.1.2 Amhara

The Amhara are one of the most politically influential and powerful groups within Ethiopia. The Amharic language is one of four national languages; used in government, education and commerce. Amharic is a Semitic language from the Afro-Asiatic language family. The Amhara are found all over the country but mainly reside in the Amhara region in central Ethiopia and in Addis Ababa (http://www.ethnologue.com/web.asp). The 2008 census reported that the Amhara population is approximately ~20,000,000. The Amhara are culturally, linguistically and genetically similar to Arab populations (De Stefano et al. 2002).

2.1.1.1.3 Anuak

The Anuak are one of the smallest ethnic groups; the 2007 census estimated that the total population size is ~86,000 (http://www.ethnologue.com/web.asp). The Anuak live in small communities of ≤500 in the Gambela region of Ethiopia, which extends into southern Sudan (Lewis and SIL International 2009). They are culturally, linguistically and historically different from the dominant ethnic groups within Ethiopia. The Anuak language is part of the Eastern Sudanic branch of the Nilo-Saharan language family.

2.1.1.1.4 Maale

The Maale are also a small ethnic group (98,114 individuals according to the 2007 census) who reside in the southwest of Ethiopia in the Omo region, southeast of Jinka (http://www.ethnologue.com/web.asp). The Maale language is from the Omotic region of the Afro-Asiatic language family.

2.1.1.1.5 Oromo

The Oromo are one of the largest ethnic groups within Ethiopia; the total number of Oromo individuals was reported as 25,448,344 in 2008. The Oromo reside predominantly in the Oromo region, west and central Ethiopia; although they are also found in Kenya (Henze 2000). The Oromo language is part of the Cushitic branch of the Afro-Asiatic language family and is widely spoken in Ethiopia.

2.1.1.2 Non-Ethiopian sub-Saharan African populations

Seven *CYP3A5* exons, and their flanking introns, were re-sequenced in eight additional sample sets from sub-Saharan Africa. The entire gene could not be re-sequenced in these individuals due to time and funding constraints.

2.1.1.2.1 Asante

The Asante are a population who reside in the Ashanti province in the south central region of Ghana (http://www.ethnologue.com/web.asp). The Asante speak three dialects of the Akan language family (Twi, Asanti and Achanti) which are all from the Atlantic-Congo branch of the Niger-Congo A language family (Lewis and SIL International 2009). The Asante re-sequenced for work presented in this thesis are all Twi speakers; of which there are approximately ~1,900,000 in Ghana (2004 Ghanaian census; http://www.ethnologue.com/web.asp).

2.1.1.2.2 Bulsa

The Bulsa are a population found in the Sandema District in north central Ghana (http://www.ethnologue.com/web.asp). Like the Asante, the Bulsa speak languages from the Atlantic-Congo branches of the Niger-Congo A language family (Lewis and SIL International 2009) and the 2003 Ghanaian census estimated the total population to be ~150,000.

2.1.1.2.3 Shewa Arabs

The Shewa Arabs re-sequenced in this thesis were collected from a number of locations in the Lake Chad region of Cameroon. Shewa Arabs are spread over multiple countries including Chad, Niger and Cameroon. The first records of Shewa Arabs date back to the 14th century and Shewa Arab ancestors are believed to have migrated from Sudan to Chad (http://www.prayway.com/unreached/peoplegroups2/1752.html). The total number of

Shewa Arabs is estimated at 1,1,39,000 (http://www.ethnologue.com/web.asp) and they all speak a form of Arabic (Lewis and SIL International 2009).

2.1.1.2.4 Mambila from Somie

The Mambila were collected from the village of Somie, Adamawa Province on the Nigerian/Cameroonian border. There are estimated to be 30,000 Mambila, the Mambila language is a Bantoid language (Lewis and SIL International 2009).

2.1.1.2.5 Congolese from Brazzaville

The Congolese samples represent a mixture of ethnic groups from Brazzaville in the Republic of Congo; hereafter called "Brazzaville". Every individual speaks a Bantu language (Niger-Congo B).

2.1.1.2.6 Chewa

The Chewa are an ethnic group who reside in west-central and south-western Malawi (Lewis and SIL International 2009); the Chewa re-sequenced for this thesis were collected from Lilongwe. The Chewa speak Chichewa, which is a Niger-Congo B (Bantu) language (http://www.ethnologue.com/web.asp) and ~7,000,000 were estimated to reside in Malawi in the 2001 census.

2.1.1.2.7 Sena from Mozambique

The Mozambican samples represent a mixture of ethnic groups collected from Sena, Mozambique; hereafter called "Sena". Every individual speaks a Bantu language (Niger-Congo B).

2.1.1.2.8 Sudanese from Kordofan

Sudanese individuals collected from Kordofan also represent a mixture of ethnic groups collected from the Kordofan Mountains, just north of Khartoum. Each of the individuals speaks an Afro-Asiatic language.

2.1.2 Additional populations genotyped for the CYP3A5 geographic survey

An additional 23 populations were genotyped as part of a geographic survey, presented in chapter 3. The groups were chosen to represent as much African genetic diversity, as possible.

2.1.2.1 Armenians

The Armenians were collected from the South of Armenia; every individual speaks Armenian (an Indo-European language) as their first language.

2.1.2.2 Anatolian Turks

The Turks were collected across East and West Anatolia; every individual speaks Turkish (an Altaic language) as their first language.

2.1.2.3 Algerians

The Algerians were collected from Port Say in the North of Algeria; every individual speaks Arabic (an Afro-Asiatic language) as their first language.

2.1.2.4 Moroccan Berbers

The Moroccan individuals were collected from Ifrane and are all from the Berber ethnic group. The Berber language is an Afro-Asiatic language.

2.1.2.5 Northern Sudanese

Sudanese individuals collected from Khartoum; hereafter referred to as Northern Sudanese individuals, are a mixture of ethnic groups. Every Northern Sudanese individual speaks an Afro-Asiatic language.

2.1.2.6 Wolof

The Wolof individuals were all collected from Dakar in Senegal. A 2007 Senegalese census reported that there are 3,976,500 Wolof speakers across the world (3,930,000 in Senegal (http://www.ethnologue.com/web.asp). The Wolof language is from the Atlantic-Congo branch of the Niger-Congo A language family.

2.1.2.7 Manjak

The Manjak individuals were all collected from the Southern region of Senegal. A 2006 Senegalese census reported that there are over 300,000 Manjak speakers across the world (105,000 in the southwest of Senegal) (http://www.ethnologue.com/web.asp). The Manjak language is from the Atlantic-Congo branch of the Niger-Congo A language family.

2.1.2.8 Kasena

The Kasena individuals were collected from the Navrongo district in the north central region of Ghana. The Kasena all speak Kasem which is from the Atlantic-Congo branch of the Niger-Congo A language family. A 2004 Ghanaian census reported that there are 130,000 Kasem speakers in Ghana.

2.1.2.9 Igbo

The Igbo were collected from the Calabar region of Nigeria. The Igbo are one of the largest groups within Nigeria itself; a 1999 Nigerian census reported that there were 18,000,000 within Nigeria itself and they are spread all over the country. The Igbo language is Igbo and is from the Atlantic-Congo region of the Niger-Congo A language family.

2.1.2.10 Kotoko

The Kotoko were all collected from a number of locations in the Lake Chad region of Cameroon. The Kotoko language is from the Chadic branch of the Afro-Asiatic language family.

2.2.2.11 Mayo Darle

Cameroonians from Mayo Darle; hereafter called "Mayo Darle", are a mixture of Cameroonian ethnic groups all collected from the Mayo Darle region of Cameroon. Every individual speaks a Niger-Congo A language.

2.1.2.12 Southern Sudanese

Sudanese individuals collected south of Khartoum; hereafter referred to as Southern Sudanese individuals, are a mixture of ethnic groups. Every individual speaks a Nilo-Saharan language.

2.1.2.13 Bantu speakers from Uganda

Ugandan individuals collected from the Ssese Islands were a mixture of ethnic groups and all are Niger-Congo B (Bantu) speakers.

2.1.2.14 Chagga

The Chagga are an ethnic group who are found in Tanzania. Chagga individuals were collected from the area immediately surrounding Mount Kilimanjaro. The Chagga speak a Niger-Congo B (Bantu) language (Lewis and SIL International 2009).

2.1.2.15 Yemeni from Hadramaut

The Yemeni individuals collected from the Hadramaut region are a mixture of ethnic groups; all of whom speak an Afro-Asiatic language as their first language.

2.1.2.16 Yemeni from Sena

The Yemeni individuals collected from the Sena region are a mixture of ethnic groups; all of whom speak an Afro-Asiatic language as their first language.

2.1.2.17 Ngoni

The Ngoni are an ethnic group from Malawi. The Ngoni were all collected from Lilongwe. Ngoni speak Nyanja which is related to the Chichewa language of the Chewa from Malawi (Lewis and SIL International 2009); most of the Ngoni speak the Chewa dialect.

2.1.2.18 Tumbuka

The Tumbuka are an ethnic group found predominantly in the north of Malawi and around the west shore of Lake Malawi, but a small number of Tumbuka reside in Zambia (http://www.ethnologue.com/web.asp). The Tumbuka individuals genotyped for this thesis were collected from Lilongwe. A 2001 Malawian census reported that the total number of Tumbuka individuals is 1,142,000; and approximately 1,000,000 reside in Malawi. The Tumbuka language is from the Atlantic-Congo branch of the Niger-Congo B (Bantu) language family.

2.1.2.19 Yao

The Yao are also an ethnic group found predominantly in the southeast region surrounding Lake Malawi and bordering Mozambique, but individuals also reside in Mozambique, Tanzania and Zambia (http://www.ethnologue.com/web.asp). The Yao individuals genotyped for this thesis were collected from Lilongwe. The global Yao population is estimated to be 1,916,000 (1,000,000 within Malawi). The Yao language is from the Atlantic-Congo branch of the Niger-Congo B (Bantu) language family.

2.1.2.20 Lomwe

The Lomwe are a Malawian ethnic group found in the southeast of the country. The numbers are estimated at 250,000 and the language is a mixed language with the Lomwe dialect of Mozambique, although it is a Niger-Congo B (Bantu) language (Lewis and SIL International 2009).

2.1.2.21 Zimbabweans from Mposi

The Zimbabwean individuals genotyped in this thesis were collected from Mposi, a village near Mberwengwa, in Zimbabwe. The Zimbabweans were a mixture of ethnic groups, although all were Niger-Congo B (Bantu) speakers.

2.1.2.22 Lemba

The Lemba are a southern African, Niger-Congo B (Bantu) speaking, tribe located in Southern Africa who claim Jewish ancestry and observed many Semitic traditions including Kosher-like dietary restrictions (http://www.freemaninstitute.com/Gallery/lemba.htm). The Lemba genotyped for this thesis were all collected from Zimbabwe.

2.1.2.23 Bantu speakers from Pretoria

South African Bantu (Niger-Congo B) speakers collected from Pretoria consisted of mixed ethnic groups including the Tswana, Zulus and Sotho.

2.1.3 Samples used for integrative analyses

The African samples were analysed in a global context by performing comparative analyses with *CYP3A5* re-sequencing and genotyping data that are available from online resources and the literature. Some individuals have been genotyped or re-sequenced as part of multiple human DNA panels and care was taken to ensure that individuals were only analysed once.

2.1.3.1 Coriell DNA samples

The bulk of integrative analyses were performed using *CYP3A5* data which have previously been published (Thompson et al. 2004; Thompson et al. 2006), and were kindly provided by Dr Emma Thompson and Professor Anna Di Rienzo from the University of Chicago. The *CYP3A5* study was performed using samples from three Human Variation panels

of the Coriell Cell Repositories (24 individuals of recent European ancestry, 23 African-Americans and 23 Han Chinese from Los Angeles). The Coriell Institute have set up a number of projects in order to examine the effect of human genetic variation, in multiple populations, on healthcare outcomes including neurological diseases and cancer risk (http://www.cogforlife.org/imr90CoriellFullReport.pdf). Coriell DNA samples are all derived from cell cultures, and since the publication of the *CYP3A5* surveys by Thompson *et al*, the number of populations in the repositories has increased (http://ccr.coriell.org/). The Europeans, Han Chinese and African-American individuals are all unrelated and healthy individuals. Unlike for the UCL collection, the Coriell samples are a mixture of male and female samples.

2.1.3.2 Human Genome Diversity Panel-Centre d'Etude du Polymorphisme Humain

Genotype data generated for the geographic survey (presented in chapter 3) were analysed in a global context using *CYP3A5*1/CYP3A5*3* genotyping data for the Human Genome Diversity Panel-Centre d'Etude Polymorphisme Humain (HGDP-CEPH), previously published (Thompson et al. 2004), and kindly provided by Dr Emma Thompson and Professor Anna Di Rienzo from the University of Chicago. The HGDP-CEPH panel is a large, and widely used, collection of DNA samples from around the world. A total of 1063 individuals from 52 global populations are available from the HGDP-CEPH collection, the populations were collected with the aim of aiding studies of sequence diversity and human migratory and population history (http://www.cephb.fr/en/hgdp/diversity.php/) and like the Coriell cell repositories, the DNA is derived from lymphoblastoid cell lines. 1028 HGDP-CEPH individuals from 51 global populations were analysed alongside the African genotyping data in chapter 3. Geographic co-ordinates were also available for all samples genotyped from the panel to enable comparisons of latitude with frequencies of low/non-expresser *CYP3A5* alleles.

2.1.3.3 NIEHS populations

The National Institute of Environmental and Health Sciences (NIEHS) have set up a project to examine the effect of environmental factors and inter-individual sequence variation on disease risk in human populations (primarily from the USA) (http://egp.gs.washington.edu/). The NIEHS SNPs programme is one part of the project in which identified candidate genes were re-sequenced to identify common and rare polymorphisms for functional analysis and population-based studies. To date *CYP3A5* has been re-sequenced in 95 individuals from five ethnic groups (12 Yoruba, 15 African-Americans, 22 Europeans, 22 Hispanics and 24 East Asians [12 Japanese and 12 Chinese]) at

25X coverage. Although NIEHS have a large re-sequencing survey, the data available online do not currently allow for the extent of missing data, per individual, to be accurately deduced and so integrative analyses were limited to genotype data alone.

2.1.3.4 HapMap samples

The International HapMap project is a collaborative initiative which aims to determine the common patterns of human genetic variation to aid researchers examine genetic variants which affect healthcare outcomes; such as the response to pharmaceutical drugs (http://hapmap.ncbi.nlm.nih.gov/). Although common haplotypes occur within all human populations, there are population specific differences in frequency, and in haplotype structure, between global populations. The HapMap project comprises a total of 1184 samples from 11 global populations (see Table 3.8) (Altshuler et al. 2010). Genotype data from the International HapMap consortium were used for integrative analyses with the data generated from the *CYP3A5* geographic survey in chapter 3. Genotype data for the *CYP3A5*1/CYP3A5*3* locus were available for each of the 11 populations; *CYP3A5*6* data were available for some of the populations; *CYP3A5*7* data are not currently available from the HapMap project.

2.1.3.5 1000 Genomes data

The 1000 Genomes project was launched in 2008 and aims to be the most detailed survey of human genetic variation in over 1000 study participants from 29 populations using next generation sequencing technologies (http://www.1000genomes.org/about). The project aims to characterise the most common genetic variants (those with a global frequency of ≥1%). The 1000 Genomes data promise to be invaluable in evolutionary and medical based studies (Kuehn 2008; McGuire 2008; Gamazon et al. 2009; Patterson 2011). However there are limitations; currently the data available for the majority of individuals re-sequenced (~1200) are low coverage (X2) (Pennisi 2010). For each individual, the genome needs to be re-sequenced in segments; which requires deep sequence coverage (estimated at X28 by the 1000 Genomes consortium) to ensure that the entire genomic sequence is covered. The project aims to complete re-sequencing for the full set of samples at 4X coverage (http://www.1000genomes.org/about). While this will not provide the complete genotype of each sample, 4X coverage is expected to allow for the majority of common variants (at a frequency of ≥1%) to be identified (http://www.1000genomes.org/about). For work presented in this thesis, 1000 Genomes re-sequenced data were not included for analysis as it is possible that a substantial amount of genetic variation would be missed in integrative analyses with the African datasets. Given that the 4X coverage re-sequencing data are due to be released by June 2012, it was not possible to incorporate higher coverage re-sequencing

data into the analysis due to time constraints. However, genotyping data for the most common functionally important *CYP3A5* variants were extracted from the website and analysed as part of the geographic survey (see chapter 3).

2.2 Experimental Methods

2.2.1 DNA extraction

DNA extraction from buccal swabs had been, previously performed, using a phenol chloroform method (Freeman et al. 2003). Briefly, 800µl of 2µg/ml proteinase K solution was added to the sample tubes; which were left to incubate at a minimum of 2 hours at 60°C. Following incubation, 800µl of sample solution was transferred to new 1.5ml eppendorf tubes containing 600µl of a 1:1 ratio of phenol/chloroform. The tubes were inverted to mix and then centrifuged at 16,000xg for 10 minutes. Following centrifugation the supernatant/aqueous phase was removed and transferred to clean 1.5ml eppendorf tubes containing 600µl of a 1:1 ratio of phenol/chloroform and 30µl of 5M sodium chloride. The samples were mixed and centrifuged again at 16,000xg for 10 minutes and the supernatant/aqueous phase again transferred into a clean 1.5ml eppendorf tube containing 700µl chloroform and centrifuged again. Following the third centrifugation the supernatant/aqueous phase was transferred to a clean 1.5ml eppendorf tube containing 700µl of 100% isopropanol solution. The tubes were inverted to mix and left to cool at -20°C for a minimum of 2 hours. Following incubation the samples were centrifuged at 16,000xg for 12 minutes after which the supernatant was poured away and the tubes inverted at an angle for one minute to allow the final remnants of supernatant to drain away. 800µl of 70% ethanol was then added to each tube followed by a centrifugation step of 16,000xg for 10 minutes. The supernatant was discarded and the tubes left to drain (by inverting at an angle) for 20 minutes. The DNA pellet was resuspended in 300µl of 1x T.E. buffer and the tubes incubated at 56°C for 10 minutes with occasional mixing. Samples were pulse centrifuged and stored at -20°C.

2.2.2 Primer design

Primers were designed by hand using a reference sequence obtained from NCBI (http://www.ncbi.nlm.nih.gov/). The *CYP3A5* reference was aligned with homologous *CYP3A* sequences: *CYP3A4, CYP3A7* and *CYP3A43*, to ensure primer specificity. The primers were then checked using the BLAT resource on UCSC Genome Browser (http://genome.ucsc.edu/) and using NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to check specificity. Primers were provided by MWG Biotech®.

2.2.3 Polymerase chain reaction (PCR)

DNA was amplified, by polymerase chain reactions (PCR), in 10µl-reaction volumes containing 50pmol of each primer, 0.2 units *Taq* DNA polymerase (HT Biotech, Cambridge, UK), 0.2µM dNTPs, 0.1µM of 10x Buffer IV [750mM Tris-HCl, pH 8.8; 200mM $(NH_4)_2SO_4$; 0.1% (v/v) Tween® 20 and 25mM $MgCl_2$] (Thermo Scientific®), and between 10-20ng genomic DNA. Cycling conditions were individually optimised for each PCR reaction; each assay included an initial denaturisation followed by a minimum of 38 cycles of denaturisation, primer annealing and elongation. Details of oligonucleotides used to amplify specific genomic regions are provided in Table 2.1. PCR cycles were performed with the following conditions: 95°C for 5 minutes, followed by 38 cycles of 95°C for 40 seconds, (optimum annealing temperature, see Table 2.1)°C for 40 seconds and 72°C for 40 seconds.

Approximately half of the *CYP3A5* region re-sequenced in Ethiopian populations was performed externally by Macrogen USA (http://www.macrogenusa.com/). Re-sequencing of the 4448 base pair region in non-Ethiopian populations, and half of the *CYP3A5* gene region in Ethiopian populations, were performed by me.

2.2.4 Gel electrophoresis

Prior to sequencing and following PCR, 2µl of the PCR reaction mixture (mixed with 2µl of loading buffer) was loaded on a 2% agarose gel (containing 80ml 1x TBE buffer, 1.6 grams of agarose and 2µl ethidium bromide), and photographed under ultraviolet light.

2.2.5 PCR clean-up

Prior to re-sequencing the amplified PCR product was "cleaned up" in order to remove excess nucleotide primers and salts. PCR clean-up was performed using a polyethylene glycol precipitation method; 30µl of Microclean (i.e. 3x the volume of the PCR reaction mixture) was added to each well and mixed thoroughly before centrifuging at 3870rpm for 60 minutes. Following centrifugation the supernatant, which contains excess nucleotides, primers and salts, was removed by removing the plate lids and inverting the plate onto a piece of tissue and centrifuging gently at 300rpm for 30 seconds. A 150µl ethanol wash (70%) was then added to each well to remove any excess nucleotides or contaminants; this time the plate was not mixed. The plate was then centrifuged again at 3870rpm for 10 minutes and the wash discarded by inversion and centrifugation as described above. The purified DNA pellets were air dried at room temperature, to allow for any excess alcohol to evaporate, and re-suspended in 10µl of distilled water.

Table 2.1: A list of the primers used for PCR amplification and sequencing of CYP3A5. Rows shaded in light blue show the regions of *CYP3A5* which were only re-sequenced in Ethiopians, boxes shaded in white show regions of *CYP3A5* which were re-sequenced in Ethiopian and North, West, West Central and South East African populations.

Region of *CYP3A5*	Fragment name	Primer sequences	Fragment size (base pairs)	Position from the ATG start codon	Position on chromosome 7 (NCBI Build 132)	Annealing temperature (°C)	Number of PCR cycles
5' upstream	U5	F: 5´-CACTTTGTTGATTGCTTTCTTTGTG-3´ R: 5´-CTGGGGGAAAAGACAGTCTCTTC-3´	631	-2500 › -1870	99280019 - 99279389	60	40
	U4	F: 5´-CATTTGTGAATAGTGGCTATTGTG-3´ R: 5´-TGAAGAACTACCCCACAAGCA-3´	632	-1973 › -1342	99279492 - 99278861	60	40
	U3	F: 5´-ACTCAAATGCAGCCACACTGTGT-3´ R: 5´-CACAATATCCAGAAATCCCCATGC-3´	633	-1449 › -817	99278968 – 99278336	55	40
Proximal promoter	U2	F: 5´-ACACATCTTTACCCACGAAATTC-3´ R: 5´-TTATGAGGAATTAAGTGGCAGAA-3´	520	-914 › -395	99278433 – 99277914	55	40
Proximal promoter	U1	F: 5´-CGCCACTTTCCTTCTTCAACTG-3´ R: 5´-TAAGGAAAAATTTTAGCAGAAGGGG-3´	511	-511 › -22	99278030 – 99277541	55	40
Exon 1		F: 5´-GAACCCAGAACCCTTGGACT-3´ R: 5´-TCCCACTACCAAATGCTGTCCCT-3´	598	-277 › 320	99277796 – 99277199	59	38
Exon 2		F: 5´-AGACTTCAGCTGCTTTCAGC-3´ R: 5´-TGGGCTACCATATCATGCACAGG-3´	595	3462 › 4056	99274057 – 99273463	61	38
Exon 3		F: 5´-AGCTTCCTTCAACTGCCAGTGAA-3´ R: 5´-ACCACAACTTTGCACAAAGGCT-3´	594	5180 › 5773	99272339 – 99271746	63	38
Exon 4		F: 5´-ATGGGCCCCACACCAACTGC-3´ R: 5´-TACCACTGGGCGGGACAGGAT-3´	715	6747 › 7461	99270772 – 99270058	64	38
Exons 5 and 6*		F: 5´-TACACTCAGAAGAGGCTAGGCA-3´ R: 5´-CATCTTACCCAATGCAAGGCAA-3´	1226	12444 › 13670	99265075 – 99263849	58	40
Exon 7		F: 5´-TATGACTGGGCTCCTTGACCT-3´ R: 5´-TTTGTGGTGGGGTGTTGACAGCT-3´	618	14324 › 14941	99263195 – 99262578	61	38
Exon 8		F: 5´-GTCGCCGGCCTGAAAGAAGGGC-3´ R: 5´-ATTCTTTACCAATCTGTGATATGA-3´	651	15641 › 16291	99261878 – 99261228	58	40
Exon 9		F: 5´-AGATGGAACCGGCAACTCTTT-3´ R: 5´-CCAAGTAGAGGTTCTCACTTGGTG-3´	691	16708 › 17398	99260811 – 99260121	58	40
Exon 10		F: 5´-TGGGAAAAAGCCTACCCCAT-3´ R: 5´-TCTCCTCAGAGGCTTCCTAC-3´	678	18861 › 19538	99258658 – 99257981	55	40
Exon 11		F: 5´-CCCTGGGGTGAGGATGGTCT-3´ R: 5´-TGTCTTGTGCTGGGACTGTGGATG-3´	671	26880 › 27550	99250639 – 99249969	61	38
Exon 12		F: 5´-TCTCATCTCAAGAAACGCTCCT-3´ R: 5´-CATGTCATGCTAATCTGTGTGGAC-3´	607	29434 › 30040	99248085 – 99247479	55	40
Exon 13		F: 5´-ACGATGGATGGTGAGTGCTT-3´ R: 5´-TCTGATGAGAGCTCAGGAGGAGTT-3´	600	31297 › 31876	99246222 – 99245643	58	40
3' downstream	D1	F: 5´-TTGCTGGTTTTCAGTCATTCAGT-3´ R: 5´-GTTATTCTAAGGATTTCTACTT-3´	522	31592 › 32113	99245927 – 99245406	55	40
	D2	F: 5´-TCTTGTCCACCTTAATGTGTGGCT-3´ R: 5´-ATTAAGCGAAGTGATAAAATCCC-3´	571	32014 › 32584	99245505 – 99244935	55	40
	D3	F: 5´-GTCTATCTATCCATCTATCTATCT-3´ R: 5´-CGTCAGTTGATTGGGCAGCATGT-3´	537	32447 › 32983	99245072 – 99244536	55	40
	D4	F: 5´-AAGTGCTACCAATTTTGTACGT-3´ R: 5´-AAGTATACTGGAAGCTAGGTGTG-3´	652	32822 › 33483	99244697 – 99244036	60	40
	D5	F: 5´-TATACTTTGAAGTCAGGTAAAT-3´ R: 5´-TGATTTCTAAATGATATTTCCAT-3´	531	33343 › 33873	99244176 – 99243646	55	40

* Only exon 6 and the flanking intron 5 and intron 6 were re-sequenced in non Ethiopian sub-Saharan African populations.

2.2.6 Sequencing

Sequencing was performed in 10µl reaction volumes; 2µl of cleaned up DNA was transferred to a new plate and heated to 95°C for 5 minutes to evaporate the excess distilled water. Following this, 2.15µl of 5x sequencing buffer (ABI Biosystems), 0.35µl BigDye v3.1 (AB Biosystems) 0.32µl of 5µM sequencing primer (see Table 2.1) and 7.18µl distilled water were added to each well. Sequencing cycling conditions were as follows: 96°C for 1 minute followed by 25 cycles of 96°C for 10 seconds, 55°C for 5 seconds and 60°C for 4 minutes.

As for PCR reactions, the sequencing assays needed to be cleaned up following the sequencing reaction to remove any contaminants. 2.5µl of 125mM EDTA was added to each well followed by a 30µl ethanol wash (100%), the plate was then mixed and centrifuged at 3870rpm for 60 minutes. Following centrifugation the plate lids were removed and the supernatant removed by inverting the plate onto a piece of tissue and centrifuging at 300rpm for 30 seconds. A final 30µl ethanol wash (70%) was then added to each well and the plate centrifuged, without mixing, at 3870rpm for 10 minutes. The supernatant was again removed by inverting the plate and centrifuging at 300rpm for 30 seconds. The plate was then left to air dry for 15 minutes at room temperature.

2.2.6.1 Sequencing strategy

All regions involved in transcription and translation of the gene were re-sequenced to identify variants which could affect protein expression. Appendix A shows the regions of the entire *CYP3A5* gene region which were re-sequenced. 2500 base pairs of genomic sequence located upstream of the ATG start codon, including the proximal promoter of *CYP3A5*, were re-sequenced. Re-sequencing of this region would identify all variants that can influence *CYP3A5* transcription, including its initiation; no enhancer is known to initiate transcription of the gene. 2500 base pairs of genomic sequence located downstream of exon 13, including the 3´ untranslated region, were also re-sequenced. Re-sequencing of this region would identify all variants which can affect *CYP3A5* mRNA stability and processing; thus influencing protein translation. A -GT microsatellite repeat was also genotyped in this region, located ~1400 base pairs downstream of exon 13; microsatellite repeat data were used as one method of dating identified variants (chapter 7).

The entire *CYP3A5* coding region; all 13 exons plus the flanking intronic sequences, were re-sequenced. Variants which occur in coding regions can potentially affect *CYP3A5* transcription, mRNA processing and protein translation. Few deleterious mutations are likely to occur, at high frequencies, in the coding regions of important genes. However, this is not

always observed in genes which are not widely expressed in certain populations (Nachman and Crowell 2000). For *CYP3A5,* a known low/non-expresser allele (*CYP3A5*3*) has risen to high frequencies in populations outside of sub-Saharan Africa (discussed in chapter 3). In these populations, the accumulation of additional null/low expresser alleles would not essentially matter and selection for removing these alleles would be weak. However if, as it is believed, *CYP3A5*3* was driven to high frequency by selection; a paucity of variation in the coding region of the gene would be seen. An examination of the structure and distribution of variation over the *CYP3A5* coding region will indicate whether selection has acted on the gene and whether the *CYP3A5*3* allele is likely to have rapidly risen to high frequency.

As indicated in Table 2.1, not all regions of *CYP3A5* were re-sequenced in each of the thirteen sample sets. The entire gene was re-sequenced in Ethiopian populations, however only exons 1, 2, 3, 4, 6, 7 and 11, plus the flanking introns, were re-sequenced in the remaining sample sets. Given time and funding constraints it was not possible to re-sequence the entire *CYP3A5* gene in individuals from North, West, West Central and South East Africa. However, the regions encompassing the *CYP3A5*1/CYP3A5*3, CYP3A5*6* and *CYP3A5*7* loci were re-sequenced, as well as part of the proximal promoter. Table 2.2 shows the *CYP3A5* regions re-sequenced in each population and provides information on the total amount of re-sequencing data obtained for each group.

2.2.7 Sequencing analysis

10µl of HiDi formamide (ABI Biosystems) was then added to the plates and denatured at 95°C for five minutes. Sequencing was performed in an ABI 96-capillary 3730xl DNA Analyser. Sequencing chromatograms were read using Sequencher version 4.7. Chromatograms were analysed, a minimum of 4x, using Sequencher version 4.7 and were aligned with a reference sequence downloaded from UCSC genome browser (hg18).

For samples re-sequenced in-house, sequencing was performed in both directions, to duplicate the genotype calls at each locus. For samples re-sequenced by Macrogen USA, all dubious and novel identified polymorphisms were checked in-house by re-sequencing.

Table 2.2: Information on the regions of CYP3A5 re-sequenced in this study, and the total amount of sequencing data generated for each sample set. Sample sizes are listed in chapters 3-5.

Country	Sample set	5´ upstream region	Proximal promoter	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Exon 8	Exon 9	Exon 10	Exon 11	Exon 12	Exon 13	3´ UTR	3´ downstream	Total amount of sequence (base pairs)
Ethiopia	Afar																		12,237
	Amhara																		12,237
	Anuak																		12,237
	Maale																		12,237
	Oromo																		12,237
Sudan	Kordofan																		4448
Ghana	Asante																		4448
	Bulsa																		4448
Cameroon	Shewa Arabs																		4448
	Mambila																		4448
Congo	Brazzaville																		4448
Malawi	Chewa																		4448
Mozambique	Sena																		4448

2.2.8 Genotyping

2.2.8.1 TaqMAN allelic discrimination

TaqMAN allelic discrimination is dependent upon the 5'-3' nuclease activity of AmpliTaq Gold DNA polymerase (Lyamichev et al. 1993), during each successive round of PCR (Figure 2.3). In addition to primers, minor groove binding probes, containing one of two possible fluorescence dyes (VIC or FAM) and a quencher (NFQ) are also used in the PCR. Separation of a reporter dye from its quencher, due to cleavage by AmpliTaq Gold DNA Polymerase, results in an increase in fluorescence which is measured. Both primer and probe must bind to their target DNA in order for amplification and cleavage to occur. Fluorescence signals are only generated if the target sequences have been amplified during PCR (McGuigan and Ralston 2002; Shen et al. 2009).

TaqMAN genotyping technology was used to genotypes samples at the *CYP3A5*1/CYP3A5*3* and *CYP3A5**6 loci. 244 samples from the Senegalese Manjak and Wolof ethnic groups, and Cameroonian samples from Mayo Darle, were genotyped at the *CYP3A5*1/CYP3A5*3* locus. 1017 samples from across the dataset were genotyped at the *CYP3A5*6* locus. Samples genotyped using TaqMAN were also re-sequenced to confirm the genotyping results.

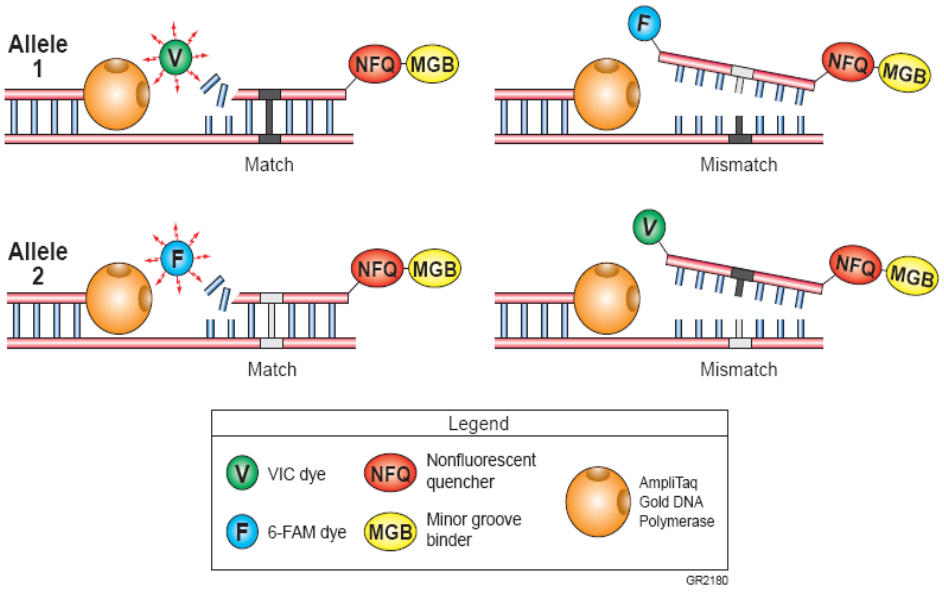
TaqMAN probes were previously designed by Applied Biosystems (*CYP3A5*1/CYP3A5*3* locus; ABiosystems, product code: C_26201809_30 and *CYP3A5*6* locus; ABiosystems, product code: C_30203959_10). The samples were diluted prior to genotyping; 2µl of 1ng/µl DNA was then dried down prior to PCR amplification. DNA was amplified in 384-well plates in 4µl-reaction volumes containing 2µl of TaqMAN Genotyping Master Mix (Applied Biosystems), 0.2µl of 20X assay mix (containing the primers and probes supplied by ABiosystems) and 1.8µl of sterile water. Polymerase chain reactions were performed with the following conditions: 95°C for 10 minutes, followed by 40 cycles of 92°C for 15 seconds and 60°C for 1 minute. Fluorescence was measured using an ABI Prism 7000 sequence detection system and genotypes were assigned with 95% confidence using ABI Prism 7000 SDS software version 2.1. No samples were genotyped at the *CYP3A5*7* locus using TaqMAN genotyping technology.

2.2.8.2 KASPar genotyping

The remaining samples were genotyped externally by KBiosciences®, UK using the KASPar method; results were viewed using SNPviewer2®. 1738 samples were genotyped at the *CYP3A5*1/CYP3A5*3* locus, 965 samples were genotyped at the *CYP3A5*6* locus, and 1982 samples were genotyped at the *CYP3A5*7* locus.

Figure 2.3: A diagram outlining the steps involved in TaqMAN allelic discrimination: taken from the ABI Biosystems manual, available at:

http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042058.pdf



The table below shows the correlation between fluorescence signals and sequences in a sample.

A substantial increase in...	Indicates...
VIC® dye fluorescence only	Homozygosity for Allele 1
FAM™ dye fluorescence only	Homozygosity for Allele 2
Both VIC and FAM fluorescence	Allele 1-Allele 2 heterozygosity

2.2.9 Microsatellite analysis

A GT microsatellite, located ~1500 base pairs downstream of the 3′ end of *CYP3A5,* was genotyped in 379 Ethiopian individuals. Microsatellite genotyping was performed using a high-throughput method adapted from (Thomas et al. 1999).

Briefly, a 456 base pair region, which has sequence overlap with the 652 base pair D4 region (see Table 2.1), was amplified using the forward primer 5′-AATATATGTGTTTGTATGTGTG-3′ and a fluorescently labelled (with FAM dye (Thomas et al. 1999)) reverse primer 5′-AAGTGCTACCAATTTTGTACGT-3′ (see Figure 2.4).

Figure 2.4: A 452 base pair region amplified for fragment analysis. The primers are highlighted in blue and the microsatellite in green.

AATATATGTGTTTGTATGTGTGagcatgcat**gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt**ataaaacatatattccctacttcactgatagggctagataacaatgacatttcaatagcaa
tgagtatacttagtgcctagatcttggcttatgaatatcattttccactgaagggaaccagagctc
gttagagaaatagctgatgccagggctaggactaaaaatgttcaagatgagcccaggacacttttt
gtgccaggaagtaagaaaacactcaaatggtttctaaatggtatttggaaatgatttctaaatgat
atttccatatgacttccaaatgatattttaaaaacctaaagactccaccaaagaactattagaact
gataaacaaattcagtaaatttgcaggatacaaaatca**ACGTACAAAATTGGTAGCACTT**

PCR amplification was performed in 10µl reaction volumes containing 0.5µM of primers, 0.2 units *Taq* DNA polymerase (HT Biotech, Cambridge, UK), 0.2µmol dNTPs, 0.1µmol reaction buffer and 0.28µl of magnesium chloride (concentration 25mM). DNA was amplified in 38 cycles, as outlined in section 2.2.2 with an optimum primer annealing temperature of 55°C (as determined during the PCR amplification stage).

Fragment analysis was performed using GS-500 ROX size standard markers (Applied Biosystems). Following PCR amplification a 10µl aliquot of a 1:90 ratio of ROX size markers: HiDi Formamide (Applied Biosystems) was added to 1.1µl of amplified PCR product. The samples were then centrifuged at 1000rpm for 1 minute and sent to Mari-Wyn Burley who performed the fragment analysis using an ABI 3730xl DNA Analyser (Applied Biosystems). Fragments were visualised using GeneMapper software (version 4.0; Applied Biosystems).

2.3 Statistical methods

Allele frequencies were calculated from genotype frequency data using the formula $frequency(A)=f(AA)+\frac{1}{2}f(Aa)$, where *AA* and *Aa* are genotypes and *A* and *a* are independent alleles at the same locus.

2.3.1 Deviations from Hardy-Weinberg equilibrium

Hardy-Weinberg Equilibrium (HWE) refers to the proportions of genotypes expected, in a randomly mating population, given allele frequency data. The expected proportion of genotypes within a randomly mating population is given by the formula:

$$\mathbf{p^2 + 2pq + q^2 = 1}$$

Where **p** is the major allele frequency and **q=1-p**. Deviations from HWE can be due to population stratification, selection or non-random mating. HWE also provides a data quality check, particularly for small sample sizes (Balding 2006). Each of the three *CYP3A5* loci were tested for deviations from HWE using Arlequin version 3.5 software (Excoffier et al. 2005).

Arlequin employs an analogue of the Fisher's Exact test when evaluating deviations from HWE (Guo and Thompson 1992). Arlequin initially generates a contingency table with the proportions of genotypes observed for each defined population in the cohort. A Markov-chain random walk algorithm is then employed to generate contingency tables, containing all possible genotype frequencies within each population, based on the allele frequency data. The *p* value of the test reflects the number of generated tables which contain genotype proportions that differ from the observed genotype frequencies. If a significance threshold of 5% is used and $p \leq 0.05$; this indicates that the observed genotypes will differ from those expected under a hypothesis of neutrality 95% of the time.

2.3.2 Chi-squared and Fisher's exact tests

Pearson's chi-squared (χ^2) tests were used to calculate inter-population heterogeneity between populations grouped by geographic region and language family in chapter 3. Pearson's χ^2 is used to compare the frequencies of one nominal variable with those obtained for another nominal variable(s) in the same population. At first a test, or χ^2, statistic is calculated using the formula:

$$X^2 = \sum_{i=1}^{n} \frac{(O - E)^2}{E}$$

Where *O* is the observed value for a class and *E* is the expected value for the population. χ^2 is the sum of the iteration of all classes. The χ^2 test measures the probability of drawing the test statistic from the χ^2 distribution based on predefined degrees of freedom (total number of classes used to calculate the χ^2 statistic minus one). The probability of drawing the value can then be compared against a pre-defined significance threshold (often 0.05): values below the threshold can be accepted as being significantly different from those expected given the values for the class.

Fisher's exact tests also compare the values for two nominal values for a given population but are more accurate than the χ^2 test when the expected values are small and are

restricted to analysing 2x2 contingency tables and two nominal variables. Both Pearson's χ^2 and Fisher's exact tests were performed using GraphPad software available freely online at http://www.graphpad.com/quickcalcs/index.cfm.

2.3.3 Haplotype inference

A haplotype is the combination of allelic states at a set of polymorphic markers on a single chromosome (Jobling et al. 2004). Haplotypes are useful in the analysis of population genetic data as they provide information on how polymorphic data in a genomic region are organised.

There are multiple methods to reconstruct haplotypes, such as pedigree analysis, physical separation of one allele from the other using laboratory techniques or by statistical algorithms for haplotype reconstruction (Jobling et al. 2004). For work presented in this thesis, pedigree information is unavailable as all individuals analysed are unrelated. Additionally, experimental methods to separate alleles on different chromosomes are expensive and laborious. All haplotype reconstruction for works presented in this thesis were performed using two statistical methods which consistently provide good estimates of haplotype frequencies from genotype data (Xu et al. 2002). All singletons, and samples with too many missing data (≥5%), were excluded from haplotype inference. All haplotype results were examined by eye for ambiguities.

2.3.3.1 A Bayesian method of haplotype reconstruction

Phase is a program which employs a Bayesian method of haplotype reconstruction from population genotype data (Stephens et al. 2001). Phase calculates the probability of two or more polymorphic sites occurring on the same or different haplotypes, based on the genotype data.

Phase uses Gibb's sampling, a Markov chain Monte Carlo algorithm, which aims to determine the posterior distribution of unknown haplotypes in a population, given the genotype data. Phase begins by resolving all unambiguous haplotypes in a population based on the genotype data. Homozygous genotype data can be resolved much more quickly than those from heterozygous loci. Once all unambiguous haplotypes have been resolved an initial probability, the prior probability, of haplotype frequencies within the sample population is guessed, P0 (Stephens and Donnelly 2003). Phase then attempts to resolve all ambiguous haplotypes within a population. Phase assumes that changes between individuals within a

population, and the population progenitor, are likely to be few. An individual with an ambiguous genotype is selected, at random, and the algorithm attempts to resolve the individual's genotype data into one of the unambiguous haplotypes. This generates a new haplotype reconstruction P1. Each successive iteration, where ambiguous haplotypes are resolved, generate a new haplotype reconstruction i.e. P0, P1, P2….Pn. Each haplotype reconstruction is a component of the Markov chain.

If an individual's genotype data resolves into two or more haplotypes of the unambiguous pool then the program assigns an equal probability of the individual having each haplotype for a single chromosome. In this case it is not possible to accurately determine the Phase of these individuals using statistical methods alone. Where an individual's genotype data cannot be resolved into any one unambiguous haplotype then the algorithm will assume that the individual has a rare haplotype. At the end of the algorithm, individuals are assigned haplotype pairs, which have been calculated to be the most probable combination given their genotype data. Singleton variants (a mutation observed in a single heterozygous individual) and individuals with data missing at least one polymorphic locus were excluded from the haplotype analysis and were assigned to haplotypes by eye, where possible.

2.3.3.2 The Expectation-Maximisation (EM) algorithm

The expectation-maximisation (EM) algorithm (Excoffier and Slatkin 1995) reconstructs haplotypes and infers haplotype frequencies which maximise the likelihood of observing the known genotype frequencies in a population. The model assumes that there are no deviations from Hardy-Weinberg equilibrium. EM estimated haplotype frequencies can vary depending upon the assumed distribution of haplotypes within a population, a so called starting point. In practice a number of different starting points are used. Each starting point generates a matrix of haplotype frequencies based on the observed genotype data. The matrix which contains haplotype frequencies which are most consistent with the observed genotype data (maximum likelihood), and do not violate the assumptions of Hardy-Weinberg equilibrium, are assumed to be correct.

For works presented in this thesis, the EM algorithm was employed using PowerMarker software (Liu and Muse 2005), which is freely available online. In order to improve the estimate of haplotype frequencies generated by EM, PowerMarker implements four starting states. The first starting point infers all possible haplotypes that could occur in a population, given the observed genotype data, and assumes that each haplotype has an equal probability of occurring. The second starting point assigns random frequencies to all possible

haplotypes that can occur in a population, given the observed genotyping data. The third starting point assumes that all haplotypes are in linkage equilibrium and the final starting point utilises the composite haplotype method.

2.3.4 Linkage disequilibrium

Linkage disequilibrium refers to the non-random association between variant alleles at two or more loci. Linkage disequilibrium was calculated using the *D´* parameter (Lewontin 1964) using LDmax software, which is part of the GOLD software package (available freely online at: http://www.sph.umich.edu/csg/abecasis/GOLD/index.html) and PowerMarker (Liu and Muse 2005).

D´ is a normalised parameter of the statistic *D*. If a haplotype is comprised of two loci, A and B, and each locus can have two alleles: i.e. at locus A, the alleles *A* and *a* can occur and at locus B, the alleles *B* and *b* can occur, then under linkage equilibrium the observed frequency of haplotype *AB* is given by the probability of the two allele frequencies of *A* and *B* i.e. $P_{AB} = P_A \times P_B$. *D* is calculated by subtracting the expected frequency of a haplotype from the observed frequency i.e. $D = P_{AB} - (P_A \times P_B)$. The alleles at the two loci are said to be in linkage disequilibrium if *D* is significantly different from zero.

However *D* is dependent on allele frequencies at two or more loci. Due to fluctuations in allele frequencies *D* is not always comparable between loci. Instead, the normalised *D´* parameter is used which improves comparability. *D´* is given by dividing *D* by its maximum possible value given the allele frequencies at the two loci. LDmax also employs a chi-square test to determine the significance of any associations between polymorphic sites at different loci.

2.3.5 Diversity comparisons

All analyses of diversity required full haplotypes of the 12,237 base pair region for each individual. Missing data were recorded as “N” and accounted for in the analyses. Full 12,237 base pair haplotypes, for each individual, were entered into Microsoft Excel 2007 and converted to FASTA format using the online programme ReadSeq (http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html). The FASTA files were then copied into notepad files. Two versions of each FASTA file were created: haplotypes were inferred excluding singleton variants from the analysis. However for calculations of analyses of departures from neutrality, singleton variants for individuals were assigned randomly on one of their two haplotypes.

All diversity analyses, and tests for departures from neutrality, were performed using DnaSP (version 5.0) (Librado and Rozas 2009) and MEGA (version 5.0) (Tamura et al. 2007).

Haplotype diversity (also called gene diversity) was measured using Nei's *h* (Nei 1987). Nei's *h* represents the proportion of choosing two different haplotypes at random from a population. Haplotype diversity calculations were performed using the programme test_h_diff and implemented in the R programming environment using code that had been written by Dr Mike Weale (Thomas et al. 2002). The multiple alleles extension of Nei's *h* was employed via the programme, along with sample bias correction. This gave the unbiased estimator of haplotype heterozygosity as:

$$\hat{h} = \frac{2N}{2N-1}(1 - \sum \hat{p}_i{}^2)$$

Where N is the number of individuals (note that 2N refers to the number of chromosomes and hence haplotypes) and $\hat{p}_i$ refers to a particular haplotype.

DNA sequence diversity was measured using the nucleotide diversity measure π (Nei 1987). π represents the weighted average number of differences, per nucleotide, between two or more haplotype sequences within a population. It measures the probability of picking two different nucleotides, at random, from a population by chance. An unbiased estimator of π is calculated using the formula:

$$\hat{\pi} = \frac{n}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{k}\hat{p}_i\hat{p}_j\hat{\pi}_{ij}$$

Where *i* and *j* are two different alleles, $\hat{p}_i$ and $\hat{p}_j$ are the frequencies of the *i*th and *j*th sequences respectively, and $\hat{\pi}_{ij}$ is the estimated nucleotide differences between the *i* and *j. K* refers to the number of different haplotypes and *n* is the number of chromosomes (Nei 1987).

2.3.6 Tests of neutrality

Tajima's *D* (Tajima 1989) compares two estimates of the expected levels of diversity within a population: θ_w; a measure of DNA sequence variation based on the observed number of segregating sites and the number of chromosomes in a sample and π; a measure of sequence variation based on the average pairwise distance between all sequences in the sample (Biswas and Akey 2006). Under neutral evolution the two values should be equal and Tajima's *D* equal to 0. In addition to calculating the value of Tajima's *D*, DnaSP calculates the

significance of deviations, of the Tajima's *D* statistic, from zero. Significantly positive values can indicate balancing selection or population subdivision, whereas significant negative values can indicate positive selection or indicate population growth (Jobling et al. 2004).

In addition to Tajima's *D,* Fu and Li's *D** and *F** (Fu and Li 1993) Fu's *FS* (Fu 1997) were calculated to test for an excess of rare variants within each of the individual populations and within the entire cohort. *D** measures the proportion of all polymorphic sites, within a population, which are singletons i.e. only observed once, *F** also measures the number of singletons and compares them to the nucleotide diversity measure, π, to ascertain how many pairwise differences between sequences within a population are attributed to rare variants (Sabeti et al. 2006). The *FS* statistic compares nucleotide diversity (π) with the number of observed haplotypes and calculates the number of haplotypes which are defined by rare variants (Sabeti et al. 2006). The significance of the number of rare haplotypes proportional to the total number observed within the population is calculated.

Fay and Wu's *H* test was performed to test for an excess of high frequency derived alleles in the cohort. Tests for a significant reduction in polymorphism levels between closely related primates and human *CYP3A5* sequences was calculated using the HKA test. A McDonald and Kreitman test was performed to detect an increase in the number of high frequency derived variants within the *CYP3A5* coding region.

2.3.7 F_{ST} and exact test of population differentiation

Both pairwise F_{ST} and Exact tests of population differentiation were calculated using Arlequin software. F_{ST} compares the total amount of genetic diversity observed within a meta-population to that seen in sub-populations. In other words F_{ST} apportions the total amount of genetic diversity observed within a population to individual sub-populations. The significance of differences in diversity observed within each sub-population is then calculated. F_{ST} is calculated by the following formula:

$$\boldsymbol{F_{ST} = (H_T - H_S)/H_T}$$

Where H_T is the total amount of variation observed in the meta-population and H_S is the estimated variation within each sub-population.

An Exact test of population differentiation (Raymond and Rousset 1995) was used to calculate pairwise genetic differences, at the haplotype level, between populations. The tests are analogous to a Fisher's Exact Test (Lee et al. 2004) in that a contingency table of frequencies are compared. However for Exact tests of population differentiation, the tables

are larger than 2x2 tables and extended to size *r* x *k*; where *r* is the number of populations being compared and *k* is the number of haplotypes. A Markov chain random walk is implemented in Arlequin to generate *r* x *k* tables based on the input data, and the probability of obtaining a table with equivalent values to those observed in the cohort is calculated. Populations are considered to be significantly different if the obtained *p* value is less than 0.05.

2.3.8 PCO analysis

Pairwise F_{ST} values were used to construct principal co-ordinates (PCO) plots to enable genetic distances between populations to be visualised in two dimensions. Principal components analysis essentially analyses a set of observations, in this case pairwise F_{ST} values, and arranges them in order of their similarity to each other. The data are initially standardised by subtracting the mean of the data; by doing this the weighted mean of the data is equal to zero. The first principal component, effectively a line of best fit, is calculated as the line that goes through the weighted mean of the data and minimises the square of the distance of each point to that line. The first principal component is calculated to capture as much of the variation between the data points as possible. The second principal component must also pass through the weighted mean but it must be completely uncorrelated to the first principal component and so pass through the data at a right angle. For the PCO plot, each axis is given an eigenvalue which corresponds to the proportion of total variation in the dataset that is captured by each PCO axis; typically the *x* axis should capture a higher proportion of the total variation. PCO plots were constructed in the R-programming environment using a code written by Dr Christopher Plaster.

2.3.9 Bioinformatics

Since the publication of the human genome (Venter et al. 2001) there has been a rapid expansion in the availability of whole genome data which have focused analyses of whole genome variation as opposed to single gene approaches alone. A number of datasets, including the genomes of numerous species, are now available online; one of the most recent examples is the ongoing 1000 Genomes project (Siva 2008; Patterson 2011). This project has expanded since it originally started and now aims to re-sequence over 2000 individuals from 28 global populations to capture as much genetic variation as possible. However for the majority of individuals that are currently re-sequenced (~1200), the depth and coverage of re-sequencing data available is low (~X2-X4) which is problematic for in-depth analysis of

rare variants (Zhang and Dolan 2010). Even analysis of coding region data, which are currently at higher coverage than non-coding data (~X4 versus X2), have high rates of type I errors (false positive results) (Tintle et al. 2011) and so the data were deemed unsuitable for integrative analyses with the data generated from re-sequencing in this thesis.

Given the rapid advances in, and increasing cost-effectiveness of, sequencing technology it has become increasingly necessary to find methods of interpreting the effects of individual sequence variants on protein and disease phenotypes. Bioinformatics analyses have aimed to provide a rapid way of interpreting both large and small amounts of re-sequencing data in a bid to understand the phenotypic implications of genetic variation. Tangible evidence of the effects of genetic variation on phenotypes can only come from experimental techniques. However bioinformatics predictions at the very least identify genomic regions, or variants, which are candidates for affecting phenotypes and so need to be examined in greater detail using experimentally established techniques. Bioinformatics analyses were performed to identify which *CYP3A5* variants are candidates for causing low/non protein expression.

2.3.9.1 ClustalW

ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/) was used to perform alignments of human *CYP3A5* with orthologous sequences in closely related primates and in the more distantly related cow and mouse (see chapter 5). The technique of aligning multiple species' sequences has been used to identify candidate evolutionarily conserved sequences (Goode et al. 2010; Lomelin et al. 2010); including identifying evolutionarily conserved regions in the promoters of different *CYP3A* genes (Thompson et al. 2004; Thompson et al. 2006). However this approach is not conclusive alone, experimental techniques will need to establish the true effect of polymorphisms within regulatory, coding and non-coding genic regions. However, it is true that functionally important regions are often conserved and so multiple sequence alignments were performed to identify such regions in the *CYP3A5* promoter, exons and exon-flanking intronic regions.

2.3.9.2 MatInspector

The regulation of transcription has a fundamental role in controlling mRNA and protein expression levels within specific cells and tissues. Promoter sequences are the main regulatory elements of gene expression; however the regulation of gene expression is often more complex and involves multiple regulatory proteins (transcription factors) and

enhancers (Alberts 2002). Differential binding of cell-type specific transcription factors which aid or hinder gene transcription can lead to tissue/cell-type specific expression of particular genes and proteins. Multiple prediction programmes, such as MatInspector, are now available which aim to detect gene regulatory motifs and predict potentially functional transcription factor binding sites (Ladunga 2010). Given the complexity of transcription factor binding *in vivo*, and the large number of identified transcription factor binding sites and regulatory elements (Ghosh 2000; Kolchanov et al. 2002) computational prediction programmes need to be able to handle deviations from set motif sequences while still being confident that a particular site is a candidate for binding of regulatory proteins in particular cells or tissues.

Analyses of regulatory motifs in the *CYP3A5* promoter were performed using MatInspector software (Cartharius et al. 2005); which examines and matches an input sequence to a library of known regulatory motifs to see if there are patterns within the input sequence which are characteristic of core transcription factor binding sites. MatInspector utilises information on real site consensus sequences, contained within the reference library, and looks for the same patterns of conserved, consensus sequences within the input sequence (of length 5-29 nucleotides). The input sequence is then scanned, and compared to a reference library, to see whether there are additional patterns consistent with known transcription factor binding motifs. The probability of the input sequence being similar to common binding motifs is then calculated. This is called a position weight matrix where the position of a particular nucleotide (of the input sequence) is compared with known regulatory motifs and the likelihood that it is a true transcription factor binding site is calculated (Lapidot et al. 2008; Ladunga 2010). MatInspector provides tissue specific information and so results can be filtered and false positives controlled by accounting for tissue and cell type specificity.

It is important to note that, like all prediction programmes, MatInspector can infer the binding potential, but not the functionality of a binding site. Tangible evidence of an effect of a mutation, or motif, on transcription factor binding can only be established through the use of experimental techniques (see section 5.4.2) (Lapidot et al. 2008; Jones and Swallow 2011).

2.3.9.3 PolyPhen2

Analysis of the effects of coding region variation on protein function was performed using PolyPhen2 software (Adzhubei et al. 2010). PolyPhen2 aligns human amino acid sequences against orthologous sequences to predict whether amino acid changes are likely to have benign, possible damaging, probably damaging, or definitely damaging effects on protein structure/function. PolyPhen2 also examines whether the mutation occurs in an

evolutionarily conserved site (as inferred from comparisons with orthologous sequences) to help predict functional consequences of an amino acid change (Adzhubei et al. 2010).

2.3.9.4 BDGP splice predictor

Almost all eukaryotic genes are composed of exonic (coding) and intronic (non-coding) segments. Initial transcription of a gene from the DNA template involves transcription of both exons and introns into the pre-mRNA transcript, in the nucleus. Intronic regions need to be excised in the nucleus prior to translation of the transcript into protein which occurs in the cytoplasm.

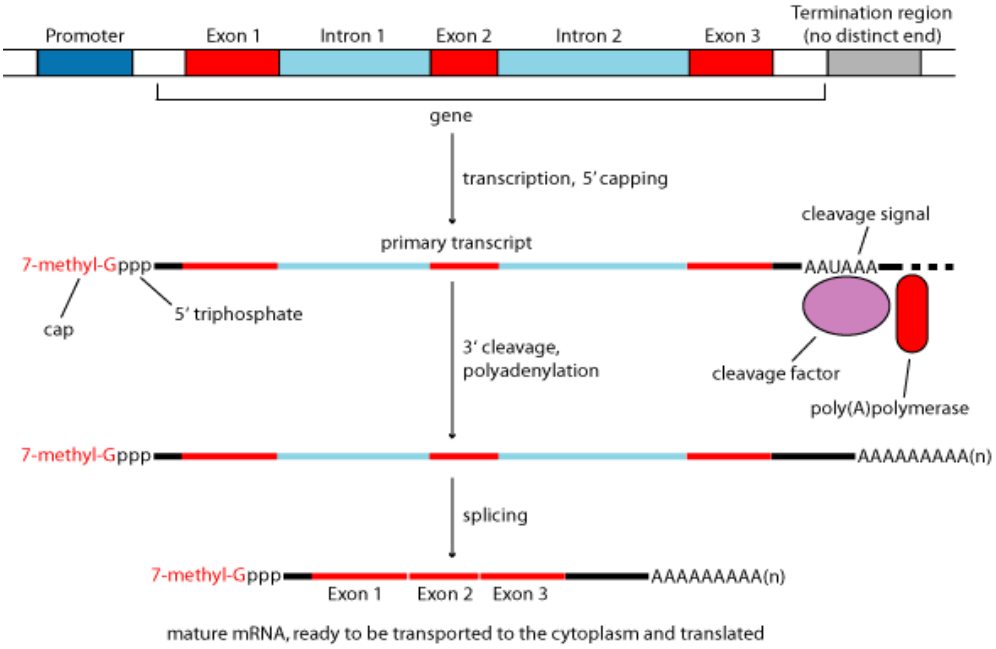
Highly conserved signals are encoded in the DNA template and recognised by nuclear proteins to enable splicing to occur (Shapiro and Senapathy 1987). Regions of genes that need to be spliced out of the transcript following transcription have two splice sites; one at the 5' end of the region, called a donor site, and one at the 3' end of the region, called an acceptor site. The consensus signal at donor sites in eukaryotic genes is GT (GU in the mRNA) and that at the acceptor site is AG (Shapiro and Senapathy 1987), additionally at the 5' splice site; a sequence of 8 nucleotides is conserved at the exon-intron boundary and a 4 nucleotide sequence was found to be conserved at the exon-intron boundary at the 3' splice site (Shapiro and Senapathy 1987). These additional signals are essential to prevent splicing of all regions of the genome which may have GT or AG repeats. Following correct splicing the mRNA transcript is ready to transport into the cytoplasm for translation (Figure 2.5).

Mutations which affect the identification of correct splice sites can occur and significantly affect protein function; certain, functionally important, variants of *CYP3A5* are defined by mutations within splice recognition regions.

Berkeley Drosophila Genome Project (BDGP) splice predictor (Reese et al. 1997) is an online programme which calculates the probability that a given polymorphism is likely to affect pre-mRNA splicing. BDGP splice predictor implements Hidden Markov Models (HMMs) to determine the likelihood that an input sequence is characteristic of a consensus splice site. The programme begins by scanning an input sequence in two rounds; one scans 15 nucleotide windows at a time searching for motifs characteristic of donor sites, as defined by (Shapiro and Senapathy 1987), and the second round scans 40 nucleotides at a time searching for motifs characteristic of acceptor sites (Shapiro and Senapathy 1987). The sequence is then scanned for consensus AG and GT splice signals to assign the likelihood that the sequence is a consensus splice site. In order to perform the analysis an ancestral nucleotide sequence is input into the software and then motifs and predictions of splice sites generated. Following this a sequence containing the polymorphism is input and the results compared with those

from the ancestral sequence to see whether there has been a change in probabilities of generating/disrupting a consensus splice site.

Figure 2.5: A diagram showing pre-mRNA processing (image has been taken from (http://xray.bmc.uu.se).



2.3.9.5 BioEdit

BioEdit is a programme which enables editing, alignment and manipulation of nucleotide and protein sequences (http://www.softpedia.com/get/Science-CAD/BioEdit.shtml). BioEdit can translate nucleotide sequences into the corresponding amino acid sequence. The software itself cannot perform any detailed analysis of sequences, however it can identify where polymorphisms can cause deviations from an ancestral amino acid sequence. All coding region (exonic) polymorphisms were analysed for their effect, if any, on amino acid sequence of CYP3A5 using BioEdit software. A comparison of the ancestral amino acid sequence with each mutant sequence was performed to determine whether an exonic polymorphism was synonymous or non-synonymous. The resulting changes were analysed further using PolyPhen2 software.

2.4 Web resources

Arlequin:	http://cmpg.unibe.ch/software/arlequin35/
DnaSP:	http://www.ub.edu/dnasp/
MEGA:	http://www.megasoftware.net/
PowerMarker:	http://statgen.ncsu.edu/powermarker/
PHASE:	http://stephenslab.uchicago.edu/software.html
LDMax:	http://www.sph.umich.edu/csg/abecasis/GOLD/docs/ldmax.html
BioEdit:	http://www.softpedia.com/get/Science-CAD/BioEdit.shtml
PolyPhen2:	http://genetics.bwh.harvard.edu/pph2/
BDGP splice:	http://www.fruitfly.org/seq_tools/splice.html
MatInspector:	http://www.genomatix.de/cgi-bin/matinspector_prof/mat_fam.pl?s=f37a3f157a0d57ed17e9c10c727382e6
UCSC:	http://genome.ucsc.edu/
NCBI:	http://genome.ucsc.edu/
Ethnologue:	http://www.ethnologue.com/web.asp
Graphpad:	http://www.graphpad.com/quickcalcs/index.cfm
TCGA:	http://www.ucl.ac.uk/tcga/software/index.html
ReadSeq:	http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html

3. The prevalence of clinically relevant *CYP3A5* alleles in Africa

3.1 Introduction

3.1.1 Previously reported frequencies of functionally important CYP3A5 variants

Four *CYP3A5* alleles are the most common determinants of interethnic variability in CYP3A5 expression and consequently affect clinical outcomes (see section 1.2.3.1). Multiple studies have reported the frequencies of the *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles in different populations. A summary of all of the currently reported frequencies, including data generated for this thesis, by population group are provided in Table 3.8. Genotype data from the 1000 Genomes genotyping databases have also been incorporated into the Table, but were not included for integrative analyses.

*CYP3A5*3* is the main *CYP3A5* variant that determines global CYP3A5 expression levels, particularly in populations outside of the African continent (Hustert et al. 2001; Lee et al. 2003; Saeki et al. 2003; Yamaori et al. 2005). In contrast *CYP3A5*6* and *CYP3A5*7* are observed primarily in populations with recent African ancestry (Hustert et al. 2001; Kuehl et al. 2001; Wojnowski et al. 2004; Roy et al. 2005; Mirghani et al. 2006; Quaranta et al. 2006). The prevalence of *CYP3A5*6* and *CYP3A5*7* in these populations may explain why CYP3A5 expression is reported to be highly variable within these populations (55-95%), despite high frequencies of *CYP3A5*1*. The Maasai from Kinyawa in Kenya, and Ethiopians have much lower frequencies of *CYP3A5*1* than other sub-Saharan African groups, and interestingly *CYP3A5*7* has not been observed in any Ethiopian individual (Gebeyehu et al. 2011). Additionally, individuals from Zimbabwe have *CYP3A5*3* allele frequencies that are comparable to those observed for non sub-Saharan African populations.

3.1.2 Rationale of study

While some African populations have been genotyped for common *CYP3A5* variants, the distribution of *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles across the continent is still largely unknown. Africa has high levels of genetic diversity (Campbell and Tishkoff 2008) and frequencies of clinically relevant *CYP3A5* alleles across the region may be highly variable. Multiple CYP3A5 drug substrates are in use across Africa; an improved understanding of common *CYP3A5* variation across diverse populations from the continent

may aid in improving healthcare outcomes within the region. Many populations within sub-Saharan Africa are prone to experiencing water shortages and it will be interesting to examine whether a previous report of an association between latitude and *CYP3A5*1* frequencies is also seen within a larger African cohort.

3.1.3 Aims of this study

The first aim of this study was to determine the frequencies of the four most common, allelic variants of *CYP3A5, CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7*, within 2538 individuals from 36 geographically and ethnically distinct populations from in and around Africa. A second aim was to use the genotypic data to infer expression patterns of CYP3A5 within the continent. A third aim was to examine whether there is a correlation between latitude and *CYP3A5*1/CYP3A5*3* allele frequencies, and CYP3A5 expression phenotypes, in Africa. A final aim was to identify potential implications of *CYP3A5* variability on medical treatment with CYP3A5 substrates, and disease risks within the continent**.**

3.2 Materials and Methods

Detailed information about all genotyped populations and statistical analyses of the data are provided in chapter 2.

3.2.1 Sample information

2538 samples were genotyped for this study (see sections 2.1.1.1.1-2.2.1.3.23). The cohort comprises 36 different sample sets, from 19 countries which span seven geographic regions, and represent six of the world's major language families. The 2538 cohort are divided into 36 distinct sample sets: some are grouped by self-declared ethnic identity and others according to where they were collected (Table 3.1).

The entire dataset overlaps with populations who were genotyped to identify the distribution of a functionally important allele of the gene encoding the drug metabolising enzyme Flavin-containing monooxygenase 2 (FMO2) (Veeramah et al. 2008) and African frequencies of clinically relevant alleles of UDP-glucuronosyltransferase 1A (*UGT1A*) genes (Horsfall et al. 2011).

1028 genotypes for 51 global populations, from the HGDP-CEPH panel, which had previously been published (Thompson et al. 2004), were combined with the 2538 sample cohort for integrative analyses (see section 2.1.1.4).

*3.2.2 Genotyping of CYP3A5*1, CYP3A5*3, CYP3A5*6 and CYP3A5*7*

Genotyping of *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5**7 was performed using a combination of Dideoxy sequencing (Slatko et al. 2001), TaqMAN allelic discrimination genotyping technology (Shen et al. 2009), and KASPar genotyping methods (http://www.kbioscience.co.uk). To ensure that results were consistent 150 samples, from across the 2538 sample cohort, were genotyped at two loci using a minimum of two methods.

3.2.3 Integrative data analyses

Distances from the equator (in kilometres) were calculated for all populations, for which geographic coordinate information was available, using the online programme: (http://www.movable-type.co.uk/scripts/latlong.html). Spearmann's Rank correlation analysis was performed, to test for a correlation between latitude and *CYP3A5* allele frequencies, in the R-programming environment and SPSS software (version 20); both gave identical results.

Given the large number of sample sets included, and statistical tests performed in this study a multiple testing correction was applied for HWE analysis, pairwise F_{ST} comparisons, exact tests of population differentiation and linkage disequilibrium analysis. The Bonferonni correction (Salkind 2007) adjusts a pre-defined significance threshold (often 0.05) by dividing it by the number of tests carried out, for example HWE analysis was carried out 36 times (once for each population genotyped) for the *CYP3A5*3* locus, therefore the initial significance threshold of 0.05 is divided by 36 to give an adjusted *p*-value of 0.00139.

Table 3.1: A list of the numbers of individuals successfully genotyped for *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* in this geographic survey. The 2538 cohort were grouped into 36 distinct sample sets: some are labelled according to self-declared ethnic identity (*a*) and others according to where the samples were collected (*b*).

Geographic region	Country	Sample set	Major language family	Latitude	Longitude	Number of individuals
Europe	Armenia	Southern Armenians (*b*)	Indo-European	40.00	45.00	100
	Turkey	Anatolian Turks (*a*)	Altaic	39.00	35.00	74
Arabian Peninsula	Yemen	Yemeni from Hadramaut (*b*)	Afro-Asiatic	14.91	48.07	82
		Yemeni from Sena and Msila (*b*)	Afro-Asiatic	16.08	49.67	37
North Africa	Algeria	Northern Algerians (*b*)	Afro-Asiatic	35.505	-1.045	161
	Morocco	Berbers (*a*)	Afro-Asiatic	34.03	-6.84	86
	Sudan	Northern Sudanese (*b*)	Afro-Asiatic	15.59	32.52	136
		Sudanese from Kordofan (*b*)	Afro-Asiatic	13.08	30.35	30
Central East Africa	Ethiopia	Afar (*a*)	Afro-Asiatic	11.602	41.360	73
		Amhara (*a*)	Afro-Asiatic	9.869	38.660	76
		Anuak (*a*)	Nilo-Saharan	7.953	34.412	76
		Maale (*a*)	Afro-Asiatic	5.715	36.643	75
		Oromo (*a*)	Afro-Asiatic	7.837	37.308	74
	Sudan	Southern Sudanese (*b*)	Nilo-Saharan	5.18	31.77	125
	Tanzania	Chagga (*a*)	Niger-Congo B	-5.38	38.05	50
	Uganda	Bantu speakers from Ssese (*b*)	Niger-Congo B	-0.57	31.45	39
West Africa	Ghana	Asante (*a*)	Niger-Congo A	5.82	-2.82	35
		Bulsa (*a*)	Niger-Congo A	10.73	-1.29	90
		Kasena (*a*)	Niger-Congo A	10.89	-1.09	47
	Senegal	Manjak (*a*)	Niger-Congo A	12.986	-15.88	94
		Wolof (*a*)	Niger-Congo A	14.687	-17.453	94
West Central Africa	Cameroon	Kotoko (*a*)	Afro-Asiatic	13.00	14.5	40
		Shewa Arabs (*a*)	Afro-Asiatic	15.05	12.11	69
		Cameroonians from Mayo Darle (*b*)	Niger-Congo A	6.47	11.55	118
		Mambila from Somie, in the Cameroonian Grassfields (*b*)	Niger-Congo A	6.00	12.5	65
	Congo	Congolese from Brazzaville (*b*)	Niger-Congo B	-4.26	15.28	55
	Nigeria	Igbo (*a*)	Niger-Congo A	4.95	8.32	88
South East Africa	Malawi	Chewa (*a*)	Niger-Congo B	-13.47	34.188	92
		Lomwe (*a*)	Niger-Congo B	-13.47	34.188	18
		Ngoni (*a*)	Niger-Congo B	-13.47	34.188	18
		Tumbuka (*a*)	Niger-Congo B	-13.47	34.188	62
		Yao (*a*)	Niger-Congo B	-13.47	34.188	56
	Mozambique	Bantu speakers from Sena (*b*)	Niger-Congo B	-17.44	35.05	85
	South Africa	Bantu speakers from Pretoria (*b*)	Niger-Congo B	-25.71	28.23	41
	Zimbabwe	Lemba (*a*)	Niger-Congo B	-23.095	29.075	24
		Zimbabweans from Mposi (*b*)	Niger-Congo B	-19.67	30.00	52

3.3 Results

*3.3.1 The distribution of CYP3A5*1, CYP3A5*3, CYP3A5*6 and CYP3A5*7 within Africa*

CYP3A5 allele frequencies are largely consistent across sub-Saharan Africa (Figure 3.1) and frequencies of the *CYP3A5*1* allele range from 33-96%. The large range reflects East African heterogeneity (Table 3.3); and differences between East Africans and other sub-Saharan Africans. North African *CYP3A5*3* frequencies range from 45-81%; the large range is influenced by high frequencies of the *CYP3A5*1* allele in the two Sudanese sample sets compared to Algerians and Moroccan Berbers. The lowest frequencies of the *CYP3A5*1* defining allele were observed in Europe (5-9%).

*CYP3A5*6* allele frequencies range from 12% to 33% in sub-Saharan Africa. *CYP3A5*6* was also observed at an average frequency of 11% in North Africa (although this is likely to be skewed due to the high frequencies observed in individuals from Kordofan in the Sudan); and 8% in the Arabian Peninsula. *CYP3A5*6* was not observed in Europe.

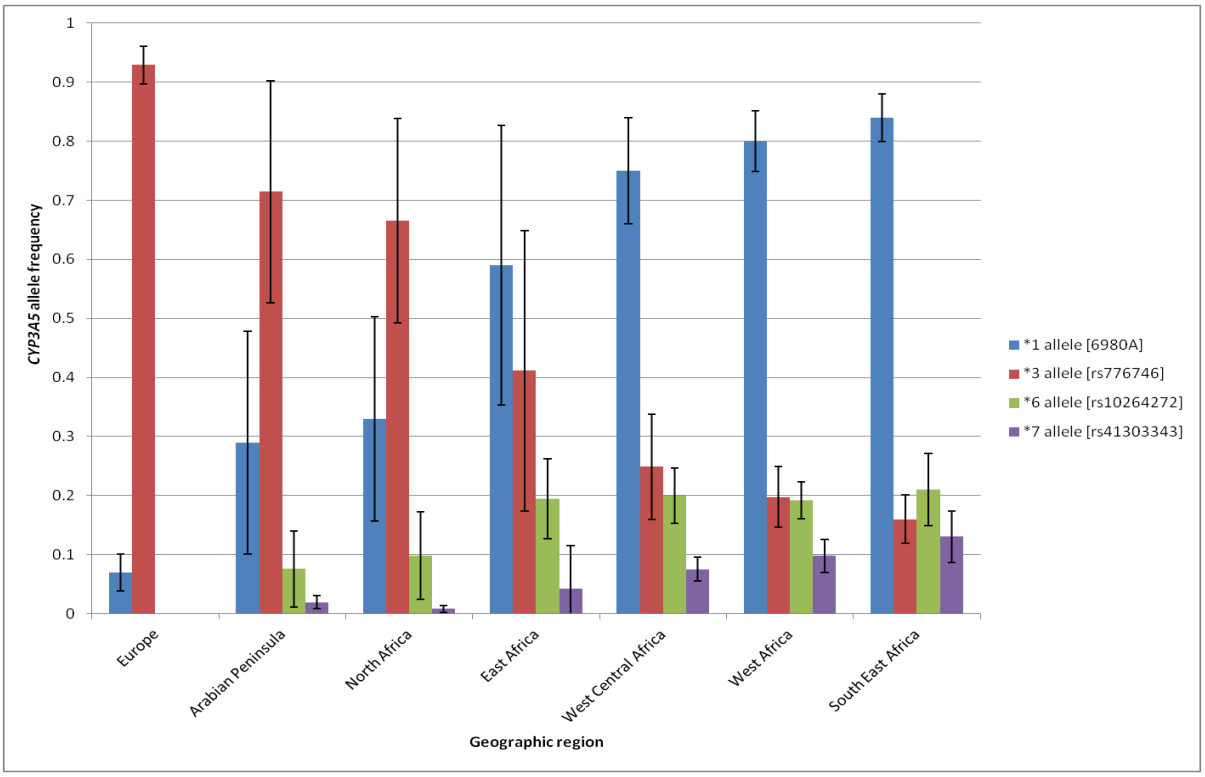
*CYP3A5*7* is largely restricted to sub-Saharan African populations, with frequencies ranging from 3-22%, from West, West Central and South East Africa. *CYP3A5*7* is also observed at low frequencies in North Africa (~1%) and in the Yemen (average frequency of 2%). Considerable heterogeneity in *CYP3A5*7* allele frequencies was observed in East Africa; Ugandans and Tanzanians had *CYP3A5*7* frequencies which are comparable to those in West, West Central and South East Africa. In contrast, *CYP3A5*7* frequencies in Ethiopia (~1%) were characteristic of non sub-Saharan African populations.

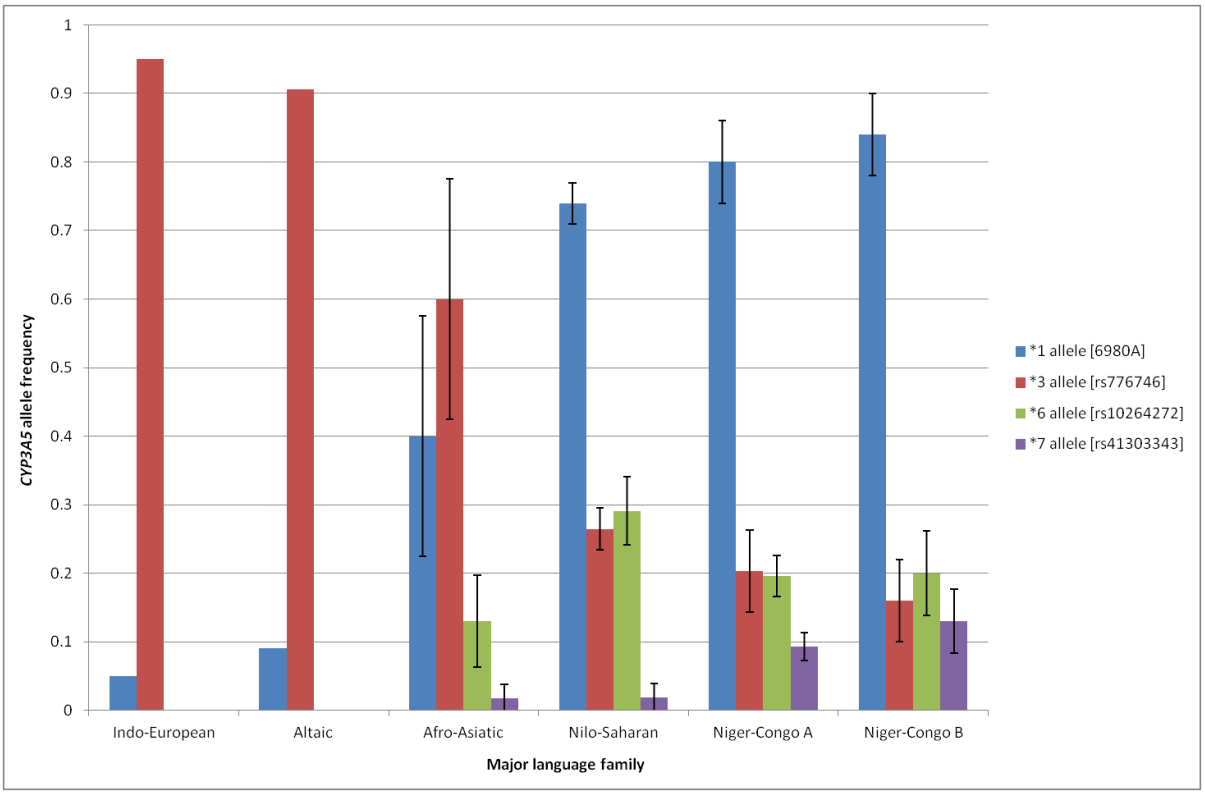
Table 3.2: A summary of the genotype and allele frequencies and χ^2 *p*-values testing for deviation from HWE at each *CYP3A5* locus by sample set genotyped for this thesis. No population deviated from Hardy-Weinberg equilibrium, following Bonferonni correction (for *CYP3A5**3: adjusted *p* value = 0.00139; correction for 36 tests, for *CYP3A5*6*: adjusted *p* value=0.0015; correction for 34 tests, for *CYP3A5*7:* adjusted *p* value=0.0017; correction for 30 tests).

Deviations from HWE cannot be calculated for monomorphic loci: labelled "N/A" on the Table. "Total" refers to the number of individuals, from a given population, successfully genotyped at each locus. Sample set refers to the grouping of individuals either by self-declared ethnicity or geography/place collected (Table 3.1).

Region	Country	Sample set	*CYP3A5*1/CYP3A5*3*						*CYP3A5*6*						*CYP3A5*7*					
			AA	AG	GG	Total	G [*3]	HWE	GG	GA	AA	Total	A [*6]	HWE	-/-	-/T	T/T	Total	T [*7]	HWE
Europe	Armenia	Southern Armenians	0	10	90	100	0.95	1.00	100	0	0	100	0.00	N/A	100	0	0	100	0.00	N/A
	Turkey	Anatolian Turks	2	10	62	74	0.91	0.11	74	0	0	74	0.00	N/A	74	0	0	74	0.00	N/A
Arabian Peninsula	Yemen	Yemeni from Hadramaut	2	21	59	82	0.85	1.00	77	5	0	82	0.03	1.00	80	2	0	82	0.01	1.00
		Yemeni from Sena and Msila	7	17	13	37	0.58	0.74	29	7	1	37	0.12	0.42	35	2	0	37	0.03	1.00
North Africa	Algeria	Northern Algerians	9	42	108	159	0.81	0.12	146	15	0	161	0.05	1.00	159	2	0	161	0.01	1.00
	Morocco	Berbers	3	28	54	85	0.80	1.00	79	7	0	86	0.04	1.00	85	1	0	86	0.01	1.00
	Sudan	Northern Sudanese	24	58	51	133	0.60	0.29	104	28	0	132	0.11	0.36	135	1	0	136	0.00	1.00
		Sudanese from Kordofan	11	11	8	30	0.45	0.16	19	10	1	30	0.20	1.00	29	1	0	30	0.02	N/A
East Africa	Ethiopia	Afar	10	31	32	73	0.65	0.61	47	26	0	73	0.18	0.11	73	0	0	73	0.00	N/A
		Amhara	14	22	40	76	0.67	0.004	55	19	2	76	0.15	0.67	76	0	0	76	0.00	N/A
		Anuak	38	32	6	76	0.29	1.00	44	25	7	76	0.26	0.23	75	1	0	76	0.01	1.00
		Maale	20	36	19	75	0.49	0.82	53	22	0	75	0.15	0.34	74	1	0	75	0.01	1.00
		Oromo	12	28	34	74	0.65	0.20	55	19	1	75	0.14	1.00	75	0	0	75	0.00	N/A
	Sudan	Southern Sudanese	74	42	9	125	0.24	0.46	58	50	15	123	0.33	0.42	117	8	0	125	0.03	1.00
	Tanzania	Chagga	28	18	4	50	0.26	0.71	36	14	0	50	0.14	0.57	41	9	0	50	0.09	1.00
	Uganda	Bantu speakers from Ssese	36	3	0	39	0.04	1.00	22	17	0	39	0.22	0.16	23	16	0	39	0.21	0.31
West Africa	Ghana	Asante	27	8	0	35	0.11	1.00	20	13	1	34	0.22	1.00	29	5	0	34	0.07	1.00
		Bulsa	58	29	3	90	0.19	1.00	61	28	0	89	0.16	0.11	69	19	2	90	0.13	0.62
		Kasena	28	17	2	47	0.22	1.00	31	16	0	47	0.17	0.32	35	12	0	47	0.13	1.00
	Senegal	Manjak	57	29	4	90	0.21	1.00	59	24	9	92	0.23	0.02	81	13	0	94	0.07	1.00
		Wolof	55	31	8	94	0.25	0.27	58	31	1	90	0.18	0.29	78	15	1	94	0.09	0.55
West Central Africa	Cameroon	Kotoko	18	21	0	39	0.27	0.04	23	16	1	40	0.23	0.65	36	4	0	40	0.05	1.00
		Shewa Arabs	26	31	12	69	0.40	0.62	42	24	3	69	0.22	1.00	60	9	0	69	0.07	1.00
		Mayo Darle	66	38	13	117	0.27	0.06	71	33	13	117	0.25	0.01	102	15	0	117	0.06	1.00
		Somie, Cameroonian Grassfields	36	28	1	65	0.23	0.16	44	19	2	65	0.18	1.00	52	13	0	65	0.10	1.00
	Congo	Congolese from Brazzaville	35	18	2	55	0.20	1.00	43	11	1	55	0.12	0.55	45	10	0	55	0.09	1.00
	Nigeria	Igbo	64	23	0	87	0.13	0.35	60	24	4	88	0.18	0.47	73	12	2	87	0.09	0.14
South East Africa	Malawi	Chewa	66	25	1	92	0.15	1.00	66	23	3	92	0.16	0.69	60	31	0	91	0.17	0.06
		Lomwe	13	4	1	18	0.17	0.39	10	8	0	18	0.22	0.53	14	4	0	18	0.11	1.00
		Ngoni	15	2	1	18	0.11	0.17	9	6	3	18	0.33	0.31	16	2	0	18	0.06	1.00
		Tumbuka	44	18	0	62	0.15	0.34	40	17	5	62	0.22	0.14	45	17	0	62	0.14	0.59
		Yao	37	18	1	56	0.18	0.67	43	12	1	56	0.13	1.00	46	10	0	56	0.09	1.00
	Mozambique	Sena	58	21	3	82	0.16	0.44	51	28	5	84	0.23	0.75	59	25	1	85	0.16	0.68
	South Africa	Bantu speakers	22	17	2	41	0.26	1.00	29	9	3	41	0.18	0.10	34	4	2	40	0.10	0.03
	Zimbabwe	Lemba	17	6	0	23	0.13	1.00	13	10	1	24	0.25	1.00	17	7	0	24	0.15	1.00
		Zimbabweans from Mposi	36	7	4	47	0.16	0.008	36	10	3	49	0.16	0.09	34	16	2	52	0.19	1.00

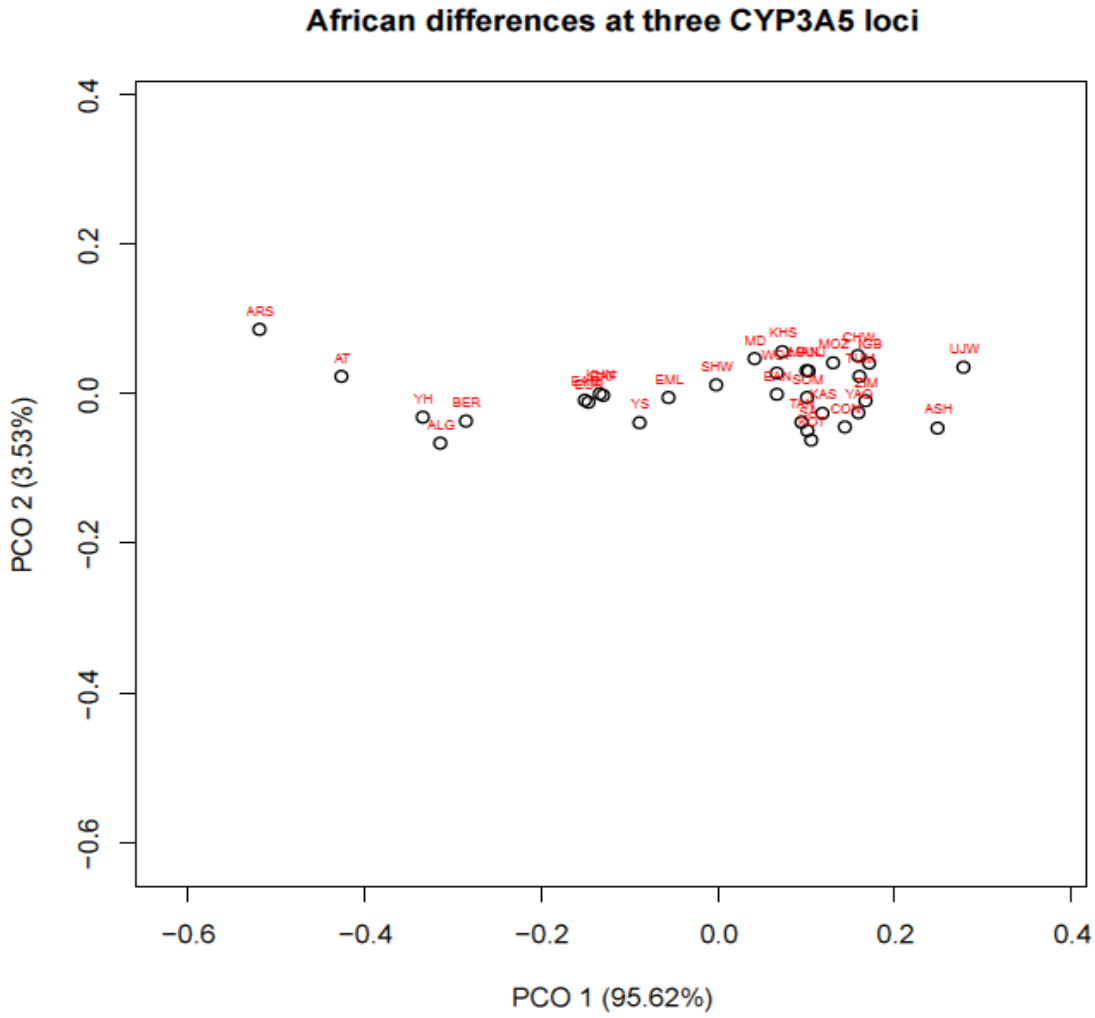
Figures 3.1 and 3.2: *CYP3A5* allele frequencies by geographic region and by major language family respectively. Error bars denote standard deviation.





Genotype data for 32 populations were used to perform pairwise F_{ST} comparisons (Supplementary Figure 1; on CD) and exact tests of population differentiation (Supplementary Figure 2; on CD). Only sample sets with 30 or more individuals were included for the analysis. Pairwise F_{ST} values were used to perform principal co-ordinates analysis (PCO) to visualise inter-population differences (Figure 3.3).

Figure 3.3: A principal co-ordinates (PCO) plot showing genetic differences between populations in which ≥30 individuals were genotyped. Supplementary Table 1 shows the associated F_{ST} and p values. Axis labels show the percentage of genetic distance captured by each axis. F_{ST} values were calculated from allele frequencies at the i) *CYP3A5*1/CYP3A5*3*, ii) *CYP3A5*6* and iii) *CYP3A5*7* loci.



2382 individuals from 32 populations were included in the analysis. Population codes correspond to the following populations: Southern Armenians (ARS), Anatolian Turks (AT), Algerians (ALG), Moroccan Berbers (BER), North Sudanese (KHN), Southern Sudanese (KHS), Yemeni from Hadramaut (YH) and from Sena (YS), Ethiopian Afar (EAF), Amhara (EAM), Anuak (EAN), Maale (EML) and Oromo (EOR), Tanzanian Chagga (TAN), Ugandan Bantu speakers (UJW), Kotoko (KOT), Shewa Arabs (SHW), Cameroonians from Mayo Darle (MD), Cameroonians from Somie (SOM), Congolese individuals from Brazzaville (CON), Nigeria Igbo (IGBO), Ghanaian Asante (ASH), Bulsa (BUL) and Kasena (KAS), Senegalese Manjak (MANJ) and Wolof (WOF), Malawian Chewa (CHW), Yao (YAO), Tumbuka (TUM), Mozambican Sena (MOZ), South African Bantu speakers (SA) and Zimbabweans (ZIM).

The results from PCO analysis show that the 32 populations broadly fall into distinct clusters. The majority of sub-Saharan African sample sets cluster together, although Ethiopians separate into two groups; the Anuak cluster with the large sub-Saharan African cohort, whereas the Afar, Amhara and Oromo cluster with North African and Yemeni sample sets. European groups are separate from all remaining populations genotyped in this study. The Yemeni from Sena are similar to the Ethiopian Anuak and Maale and Southern Sudanese individuals; which is likely to be influenced by appreciable *CYP3A5*6* and *CYP3A5*7* frequencies observed in this sample set (Table 3.2).

The most striking result from PCO analysis is that over 95% of population differentiation is explained by the first principal component. This is a result of differences in *CYP3A5*3* allele frequencies which explains the clustering of groups with similar frequencies of the allele. The second principal component differentiates groups on the basis of *CYP3A5*6* and *CYP3A5*7* allele frequencies.

The results from PCO analysis (Figure 3.3) suggest that groups from East Africa, North Africa and the Arabian Peninsula are the most heterogeneous. These results were confirmed by Pearson's chi squared tests which examined overall heterogeneity, based on data for all three *CYP3A5* loci by geographic region (Table 3.3) and major language family (Table 3.4). Only Anatolian Turks were Altaic speakers and Southern Armenians Indo-European speakers and so these language groups were excluded from this analysis. Comparisons of sample sets by major language family found significant heterogeneity only between Afro-Asiatic speaking groups (consistent with Figure 3.2). The Afro-Asiatic speakers are distributed over a wide geographic area: both within and outside of sub-Saharan Africa and it is likely that differences observed between these groups are due to geographic distance and differences in ancestry.

Table 3.3: Pearson's chi-squared test of overall heterogeneity within the seven geographic regions represented by the dataset. *P*-values which are significant at the 5% level ($p<0.05$) are shown in bold and highlighted in green.

Geographic region	Total number of individuals	Degrees of freedom	Chi-square	*P*-value
Europe	174	3	1.32	0.251
North Africa	441	9	37.61	**0.0137**
Arabian Peninsula	131	3	13.16	**0.0043**
East Africa	732	21	157.69	**5.95×10^{-23}**
West Africa	458	12	6.687	1
West Central Africa	553	15	20.69	0.133
South East Africa	588	24	13.24	0.967

Table 3.4: Pearson's chi-squared test of overall heterogeneity within the six major language families represented by the dataset. *P*-values which are significant at the 5% level ($p<0.05$) are shown in bold and highlighted in green.

Language family	Total number of individuals	Degrees of freedom	Chi-square	*P*-value
Afro-Asiatic	940	33	144.2	**1.44×10^{-16}**
Niger-Congo A	631	21	15.81	0.754
Niger-Congo B	592	33	24.45	0.819
Nilo-Saharan	201	3	1.608	0.50

Both Table 3.3 and 3.4 are consistent with Figures 3.1 and 3.2 respectively.

*3.3.2 Examining the statistical associations between CYP3A5*1, CYP3A5*3, CYP3A5*6 and CYP3A5*7 alleles*

3.3.2.1 Linkage disequilibrium

The *D´* measure of linkage disequilibrium (LD) between the three genotyped *CYP3A5* loci (see section 2.3.4) was calculated using LDmax and PowerMarker software programs, which gave identical results. *D´* was calculated using genotype data from sample sets. Individuals with missing data at any locus were excluded from this analysis. The *CYP3A5*6* and *CYP3A5*7* alleles are absent in Europe meaning that LD could not be calculated for these populations. *CYP3A5*7* is monomorphic in a number of sample sets and LD for this locus was not calculated in every population. A total of 2465 individuals were included for this analysis. Population specific *D´* values are presented in Table 3.5.

The *CYP3A5*1/CYP3A5*3* locus is located 7704 nucleotides (nt) from the *CYP3A5*6* locus and 20,145 nt from the *CYP3A5*7* locus. The *CYP3A5*6* locus is located 12,441 nt from *CYP3A5*7*. The *D´* values for the combined cohort of 2465 individuals show that LD between the three loci is high; the *CYP3A5*7* defining T insertion is in complete LD with the *CYP3A5*1* allele i.e. $D´ = 1$, $p<0.0001$; the *CYP3A5*6* defining allele is in high LD with the *CYP3A5*1* allele $D´= 0.96$, $p<0.0001$; LD between the *CYP3A5*6* and *CYP3A5*7* loci is high $D´=1.00$, $p<0.0001$. However, haplotype analysis (see section 3.3.2.2) found that the *CYP3A5*6* and *CYP3A5*7* defining alleles was non-existent; the alleles were not observed on the same chromosome. Therefore the observation that $D´=1.00$ between these two loci is between the ancestral alleles; not the derived, clinically relevant mutations. *D´* values for individual sample sets were consistent with those reported for the 2465 sample cohort.

Table 3.5: Pairwise *D´* values for each allelic combination at the *CYP3A5*1/CYP3A5*3, CYP3A5*6* and *CYP3A5*7* loci, shown by geographic region, country and ethnic group. Statistically significant *D´* values, following a Bonferonni correction for multiple testing are highlighted in green. Adjusted significance thresholds for each pairwise comparison are shown in brackets after the allelic combination. "–" indicates that at least one of the two loci is monomorphic and so LD could not be calculated.

Geography key: AP; Arabian Peninsula, NA; North Africa, EA; East Africa, WA; West Africa, WCA; West Central Africa; SEA; South East Africa

			D´ values		
Geographic region	**Country**	**Population**	**rs776746/ rs10264272** (*p*≤0.00147)	**rs776746/ rs41303343** (*p*≤0.00167)	**rs10264272/ rs41303343** (*p*≤0.00161)
Europe	Armenia	Southern Armenians	-	-	-
Europe	Turkey	Anatolian Turks	-	-	-
AP	Yemen	Hadramaut	**0.73**	**1.00**	0.47
AP	Yemen	Sena and Msila	**1.00**	1.00	1.00
NA	Algeria	Northern	1.00	1.00	1.00
NA	Morocco	Berbers	**1.00**	**1.00**	1.00
NA	Sudan	Northern	**1.00**	-	-
NA	Sudan	Kordofan	**1.00**	1.00	1.00
EA	Ethiopia	Afar	**1.00**	-	-
EA	Ethiopia	Amhara	**0.77**	-	-
EA	Ethiopia	Anuak	**1.00**	1.00	1.00
EA	Ethiopia	Maale	**0.86**	1.00	1.00
EA	Ethiopia	Oromo	**0.90**	-	-
EA	Sudan	Southern	**1.00**	1.00	1.00
EA	Tanzania	Chagga	1.00	1.00	1.00
EA	Uganda	Bantu speakers from Ssese	0.71	1.00	1.00
WA	Ghana	Asante	0.27	1.00	1.00
WA	Ghana	Bulsa	1.00	0.81	1.00
WA	Ghana	Kasena	1.00	1.00	0.07
WA	Senegal	Manjak	0.60	1.00	1.00
WA	Senegal	Wolof	**1.00**	1.00	0.12
WCA	Cameroon	Kotoko	1.00	0.26	1.00
WCA	Cameroon	Shewa Arabs	**0.85**	1.00	1.00
WCA	Cameroon	Mayo Darle	**1.00**	1.00	1.00
WCA	Cameroon	Mambila from Somie	**1.00**	1.00	1.00
WCA	Congo	Brazzaville	0.85	1.00	1.00
WCA	Nigeria	Igbo	1.00	0.74	1.00
SEA	Malawi	Chewa	1.00	1.00	1.00
SEA	Malawi	Lomwe	0.14	1.00	1.00
SEA	Malawi	Ngoni	1.00	1.00	1.00
SEA	Malawi	Tumbuka	1.00	0.30	1.00
SEA	Malawi	Yao	0.69	0.49	1.00
SEA	Mozambique	Sena	1.00	1.00	1.00
SEA	South Africa	Bantu speakers	1.00	1.00	1.00
SEA	Zimbabwe	Lemba	1.00	1.00	1.00
SEA	Zimbabwe	Mposi	1.00	1.00	1.00

3.3.2.2 Haplotype inference

Haplotypes were inferred from genotype data for 2465 individuals successfully genotyped at each of the three *CYP3A5* loci. Eight haplotypes are possible from the genotyping data (see Table 3.6) but only five were inferred from the cohort.

Table 3.6: The eight potential haplotypes which can occur based on allelic data for the *CYP3A5*1/CYP3A5*3, CYP3A5*6* and *CYP3A5*7* loci. Allelic combinations which were inferred from the cohort are highlighted in green.

Allelic combinations at all three *CYP3A5* loci			Haplotype name
CYP3A5*1/*3	***CYP3A5*6***	***CYP3A5*7***	
A	G	-	**1*
A	G	insT	**7*
A	A	-	**6*
A	A	insT	**6/*7*
G	G	-	**3*
G	G	insT	**3/*7*
G	A	-	**3/*6*
G	A	insT	**3/*6/*7*

The three low/non-expresser alleles occur on independent haplotype backgrounds indicating that they act independently on CYP3A5 expression. A low frequency recombinant *CYP3A5*3/CYP3A5*6* haplotype was inferred for 10 heterozygous individuals; which explains why *D´* between the *CYP3A5*1* and *CYP3A5*6* alleles is not equal to 1. *CYP3A5*7* always occurs on the same haplotype background as *CYP3A5*1;* consistent with the results for LD analysis (section 3.3.2.1). The results show that genotyping of the *CYP3A5*1/CYP3A5*3* locus alone does not account for additional low/non-functional alleles, in certain populations, which are in high LD with the *CYP3A5*1* allele.

Inter-population differences in gene diversity are presented in Figure 3.5. Gene diversity estimates are largely consistent across sub-Saharan Africa, although the Afar, Amhara and Oromo have lower estimates of gene diversity than other populations from the sub-continent. The highest frequencies of the AG- (*CYP3A5*1*) "expresser" haplotype were observed in sub-Saharan African populations, especially in Niger-Congo speaking groups (Figure 3.4 and Table 3.7). The distribution of the AA- (*CYP3A5*6*) haplotype was similar across all sub-Saharan African groups, in contrast the AGinsT (*CYP3A5*7*) haplotype was almost exclusively observed in Niger-Congo speakers. The recombinant GA- (*CYP3A5*3/CYP3A5*6*) haplotype was not restricted to a particular geographic region, although this may be due to its low frequency in the dataset. *CYP3A5*3* and *CYP3A5*1* have a global distribution; and frequencies of the *CYP3A5*3* haplotype are approaching 100% in Europe.

Table 3.7: The proportion of each inferred *CYP3A5* haplotype by sample set out of all haplotypes observed in each population.

Geographic region	Language family	Sample set	Frequencies of inferred haplotypes				
			AG- (**1*)	AGinsT (**7*)	AA- (**6*)	GG- (**3*)	GA- (**3/*6*)
Europe	Indo-European	Southern Armenians	0.05	-	-	0.95	-
	Altaic	Anatolian Turks	0.09	-	-	0.91	-
Arabian Peninsula	Afro-Asiatic	Yemeni from Hadramaut	0.12	0.01	0.02	0.84	0.006
	Afro-Asiatic	Yemeni from Sena and Msila	0.27	0.03	0.12	0.58	-
North Africa	Afro-Asiatic	Northern Algerians	0.14	0.01	0.05	0.81	-
	Afro-Asiatic	Berbers	0.15	0.01	0.04	0.80	-
	Afro-Asiatic	Northern Sudanese	0.27	0.003	0.11	0.62	-
	Afro-Asiatic	Sudanese from Kordofan	0.33	0.02	0.20	0.45	-
East Africa	Afro-Asiatic	Afar	0.17	-	0.18	0.65	-
	Afro-Asiatic	Amhara	0.20	-	0.13	0.65	-
	Nilo-Saharan	Anuak	0.45	0.007	0.26	0.29	-
	Afro-Asiatic	Maale	0.36	0.007	0.14	0.49	0.007
	Afro-Asiatic	Oromo	0.22	-	0.13	0.64	0.007
	Nilo-Saharan	Southern Sudanese	0.40	0.03	0.33	0.24	-
	Niger-Congo B	Chagga	0.51	0.09	0.14	0.26	-
	Niger-Congo B	Bantu speakers from Ssese	0.54	0.21	0.22	0.04	-
West Africa	Niger-Congo A	Asante	0.60	0.07	0.22	0.10	-
	Niger-Congo A	Bulsa	0.52	0.12	0.16	0.20	-
	Niger-Congo A	Kasena	0.48	0.13	0.17	0.22	-
	Niger-Congo A	Manjak	0.49	0.07	0.23	0.20	0.01
	Niger-Congo A	Wolof	0.47	0.09	0.18	0.25	-
West Central Africa	Afro-Asiatic	Kotoko	0.46	0.05	0.22	0.27	-
	Afro-Asiatic	Shewa Arabs	0.33	0.07	0.21	0.39	0.007
	Niger-Congo A	Mayo Darle	0.40	0.07	0.25	0.28	-
	Niger-Congo A	Mambila from Somie	0.49	0.23	0.18	0.10	-
	Niger-Congo B	Congolese from Brazzaville	0.59	0.09	0.12	0.20	-
	Niger-Congo A	Igbo	0.59	0.09	0.18	0.13	-
South East Africa	Niger-Congo B	Chewa	0.54	0.18	0.15	0.13	-
	Niger-Congo B	Lomwe	0.50	0.11	0.22	0.17	-
	Niger-Congo B	Ngoni	0.50	0.11	0.22	0.17	-
	Niger-Congo B	Tumbuka	0.50	0.14	0.22	0.15	-
	Niger-Congo B	Yao	0.61	0.18	0.13	0.09	-
	Niger-Congo B	Sena	0.46	0.17	0.21	0.16	0.006
	Niger-Congo B	Bantu speakers	0.45	0.10	0.19	0.26	-
	Niger-Congo B	Lemba	0.48	0.15	0.24	0.13	-
	Niger-Congo B	Zimbabweans from Mposi	0.45	0.22	0.17	0.16	-

Figure 3.4: A map showing the distribution of the five inferred *CYP3A5* haplotypes across the geographic region represented by the 36 populations in the dataset. The size of each circle is proportional to the number of individuals sampled from a given population. The allele combinations listed in the key are equivalent to those discussed in section 3.3.4.

N.B. the recombinant *CYP3A5* haplotype 5: **3/*6,* is observed at low frequency in the dataset.

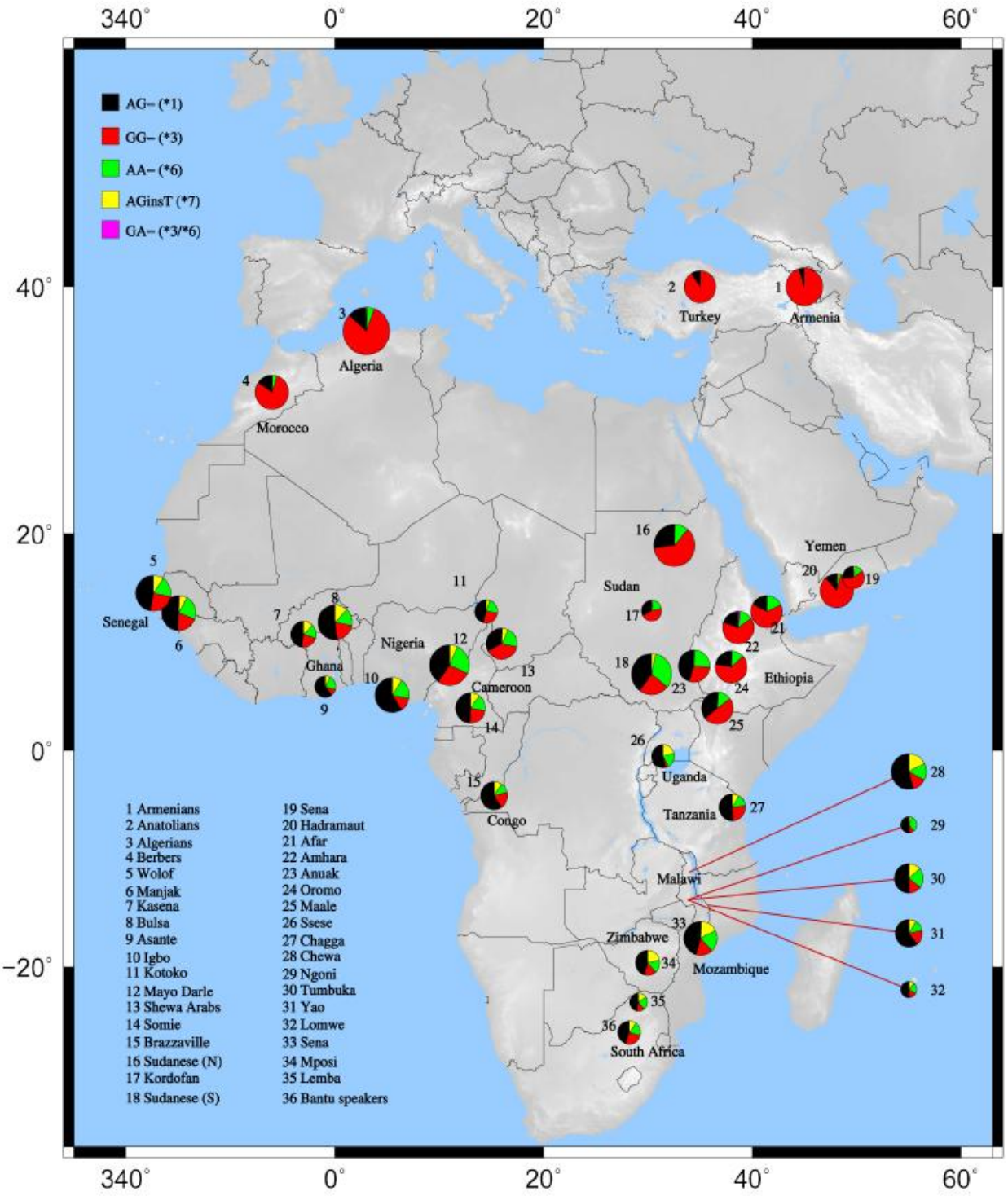
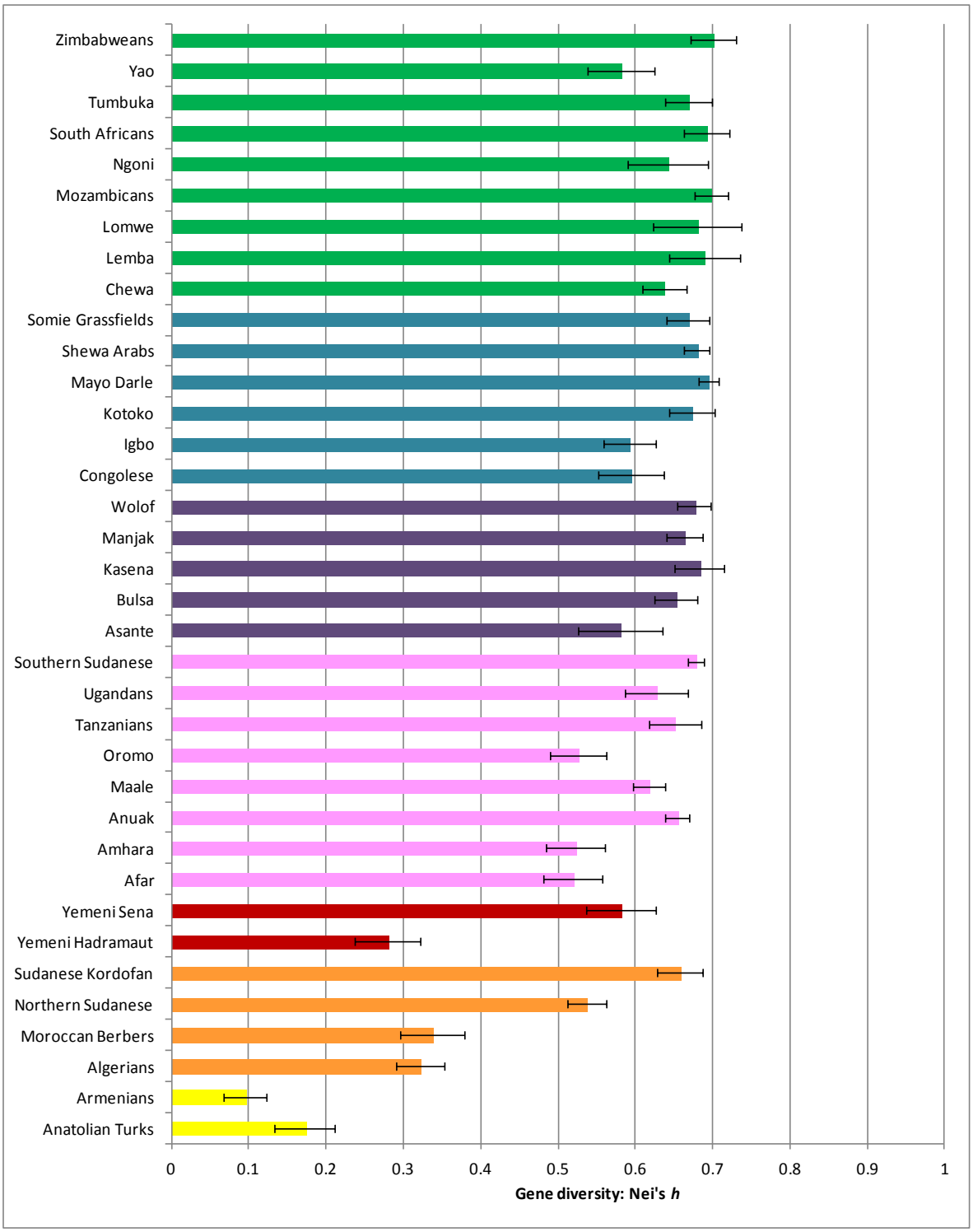


Figure 3.5: Gene diversity (Nei's h) for the five *CYP3A5* haplotypes, inferred from genotype data for each of 36 populations, for the three loci defining the *CYP3A5* alleles: *CYP3A5*1/CYP3A5*3, CYP3A5*6* and *CYP3A5*7*. Populations are coloured to the geographic region to which they belong, see Table 3.1. Error bars denote standard deviation.

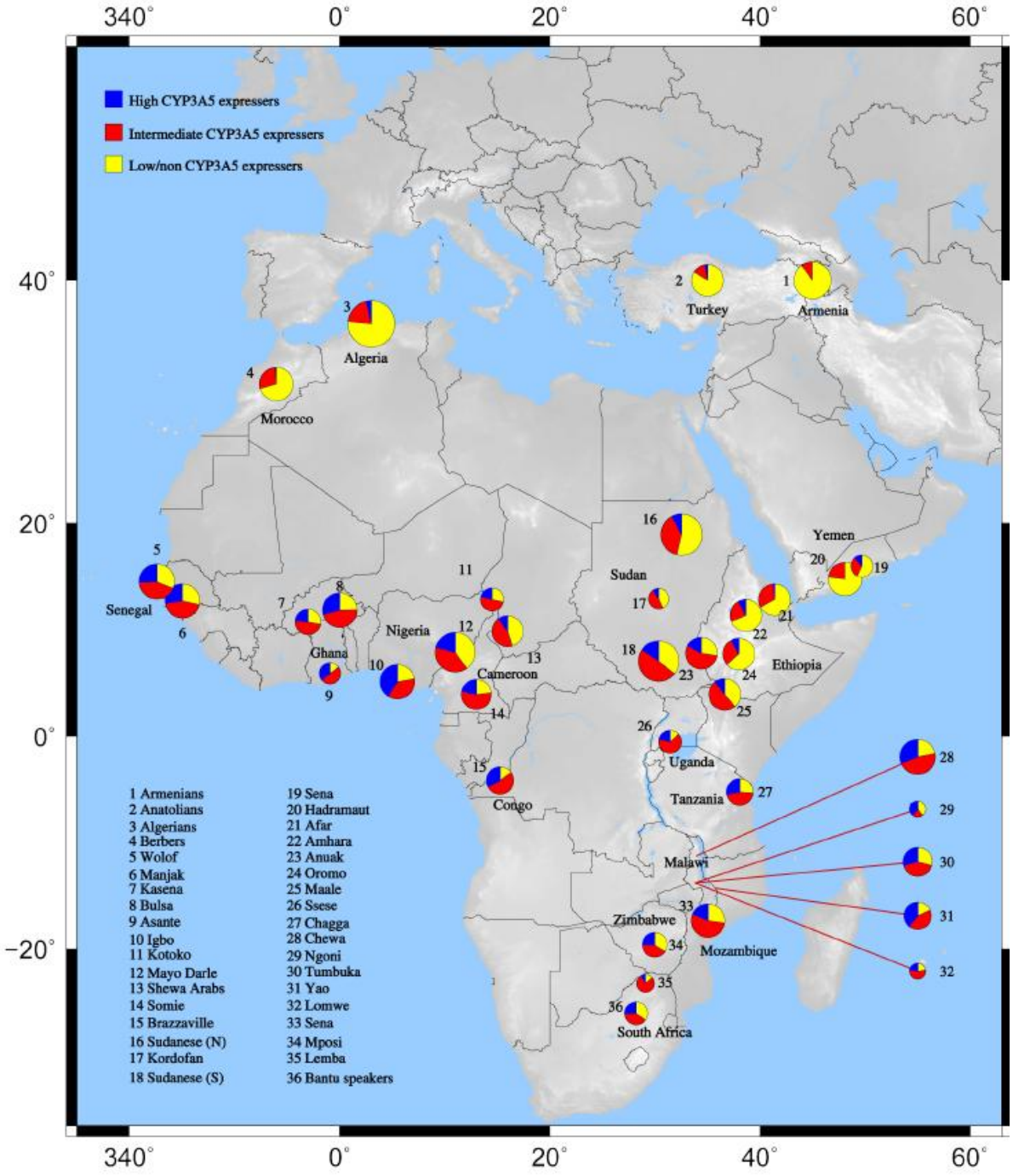


3.3.3 Inferred CYP3A5 protein expression patterns across sub-Saharan Africa

The geographic and ethnic distributions of inferred low/non-, intermediate- and high-expression phenotypes, based on individual diplotypes, are presented in Figure 3.6. Within Africa populations from South East, West and West Central Africa (all Niger-Congo speakers) are predicted to have the highest frequencies of high/normal CYP3A5 expressers. This is assuming that there are no additional variants within these populations which can affect protein expression. Populations from Europe are predicted to have the highest frequencies of individuals who are low/non-CYP3A5 expressers; consistent with previous reports (Quaranta et al. 2006). Consistent with all previous analyses, there is considerable heterogeneity in East and West Central Africa; similarities in inferred expression profiles are correlated with major language family.

The low/non-expresser and expresser haplotypes were considered as two alleles at an "expression" locus. Individuals who had two "low/non-expresser" haplotypes were assigned the pseudo genotype LL (low/low), individuals who were heterozygous for an expresser and a low/non-expresser haplotype were assigned HL (high/low) and individuals who have two copies of the expresser haplotype were assigned HH (high/high). These pseudo genotypes were used to examine whether any population deviated significantly from Hardy-Weinberg equilibrium (HWE) in their CYP3A5 expression phenotypes. No population deviated from HWE following Bonferonni correction (adjusted $p \leq 0.00139$).

Figure 3.6: A map showing the distribution of predicted CYP3A5 protein expression levels across the geographic region represented by the dataset. The size of each circle is proportional to the number of individuals sampled from a given population.



3.3.4 Examining African CYP3A5 allele frequencies in a global context

A summary of global *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* allele and genotype frequencies, including those reported in this study, are provided in Table 3.8. Previous reports have focused mainly on the *CYP3A5*1/CYP3A5*3* locus, therefore the main comparisons between the data reported in this study and global populations are for *CYP3A5*1/CYP3A5*3* data. Figure 3.7 shows the average *CYP3A5* allele frequencies by geographic region.

*CYP3A5*3* has been observed in all population groups that have been studied. The highest frequencies of *CYP3A5*3* have been observed in individuals with recent European ancestry from North America and Europe (Table 3.8) (Dally et al. 2004; Roy et al. 2005; Quaranta et al. 2006). Frequencies of *CYP3A5*3* in populations from Asia, Oceania and South America are intermediate between those in Africa and Europe (Balram et al. 2003; Quaranta et al. 2006; Diczfalusy et al. 2008; Sinues et al. 2008). Outside of sub-Saharan Africa, the largest variation in *CYP3A5*3* frequencies were observed in North America; *CYP3A5* allele frequencies vary between African-Americans and other groups from the region. Caucasian and Asian populations from the region are largely homogeneous in their allele frequencies.

Comparisons of *CYP3A5*6* allele frequencies reported in this study with those from HapMap and NIEHS populations and published data confirm that *CYP3A5*6* frequencies are highest in sub-Saharan Africa. *CYP3A5*6* frequencies are largely consistent across populations from the sub-continent, unlike *CYP3A5*1, CYP3A5*3* and *CYP3A5*7*. *CYP3A5*6* is absent in all East Asian populations, in almost every European population (a single *CYP3A5*1/CYP3A5*6* heterozygote from Tuscany, Italy was observed), and is observed at low frequencies in North Africa. The largest inter-regional variation in *CYP3A5*6* allele frequencies was observed in West Central Africa; which is almost entirely influenced by higher frequencies of *CYP3A5*6* in populations from Lake Chad comparative to other West Central Africans.

*CYP3A5*7* has not been genotyped in HapMap populations to date. Comparisons of *CYP3A5*7* frequencies with NIEHS populations and with published data confirm that the highest frequencies of *CYP3A5*7* are in sub-Saharan Africa. Appreciable frequencies of this allele were observed in ethnic Koreans (Diczfalusy et al. 2008) and North African populations (own data). *CYP3A5*7* is not present in Europe. North American populations have considerable inter-regional variation in *CYP3A5*7* allele frequencies; which is almost certainly due to inter-ethnic differences between populations sampled from the region.

Table 3.8: A table summarising all known allele frequencies of *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7*, in different population groups. Information has been compiled from reports in the literature and online databases including The International HapMap Consortium (http://hapmap.ncbi.nlm.nih.gov/) and NIEHS SNPs (http://egp.gs.washington.edu/). Care has been taken to ensure that individuals have not been counted more than once. Where the same individuals have been genotyped, the average of each allele frequency has been reported. Data generated for this thesis has also been added into the Table. A reference list is provided beneath the Table. "-" indicates missing information.

							CYP3A5*1/CYP3A5*3					***CYP3A5*6***					***CYP3A5*7***					
							Genotypes			**Alleles**		**Genotypes**			**Alleles**			**Genotypes**		**Alleles**		
Geographic region	**Country**	**Sample set**	**Longitude**	**Latitude**	**Distance from equator (km)**	**Sample size**	**AA**	**AG**	**GG**	**A (*1)**	**G (*3)**	**GG**	**GA**	**AA**	**G (-)**	**A (*6)**	**-/-**	**-/insT**	**insT/insT**	**- (-)**	**T (*7)**	**REF**
Arabian Peninsula	Yemen	Hadramaut	48.07	14.91	1658	82	2	21	59	0.15	0.85	77	5	0	0.97	0.03	80	2	0	0.99	0.01	1
		Sena	49.67	16.08	1788	37	7	17	13	0.42	0.58	29	7	1	0.88	0.12	35	2	0	0.97	0.03	1
East Asia	Numerous	East Asian	-	-	-	24	2	11	11	0.31	0.69	0	0	23	1.00	0.00	24	0	0	1.00	0	4
	Cambodia	Khmer	12N	105E	1334	11	1	4	6	0.27	0.73	-	-	-	-	-	-	-	-	-	-	2
	China	Chinese	-	-	-	108	9	35	64	0.25	0.75	108	0	0	1.00	0.00	-	-	-	-	-	5
		Dai	21N	100E	2335	10	1	7	2	0.45	0.55	-	-	-	-	-	-	-	-	-	-	2
		Daur	48-49N	124E	5393	10	0	3	7	0.15	0.85	-	-	-	-	-	-	-	-	-	-	2
		Han	26-39N	108-120E	3614	44	2	18	24	0.25	0.75	-	-	-	-	-	-	-	-	-	-	2
		Han Chinese in Beijing	-	-	-	124	12	46	66	0.28	0.72	125	0	0	1.00	0.00	-	-	-	-	-	3
		Southern Han Chinese	-	-	-	100	24	46	30	0.47	0.53	100	0	0	1.00	0.00	-	-	-	-	-	18
		Hezhen	47-48N	132-135E	5282	10	0	3	7	0.15	0.85	-	-	-	-	-	-	-	-	-	-	2
		Lahu	22N	100E	2446	10	0	5	5	0.25	0.75	-	-	-	-	-	-	-	-	-	-	2
		Miaozu	28N	109E	3113	10	1	5	4	0.35	0.65	-	-	-	-	-	-	-	-	-	-	2
		Mongola	48-49N	118-120E	5393	10	0	7	3	0.35	0.65	-	-	-	-	-	-	-	-	-	-	2
		Naxi	26N	100E	2891	9	0	5	4	0.28	0.72	-	-	-	-	-	-	-	-	-	-	2
		Orogen	48-53N	122-131E	5615	10	0	2	8	0.10	0.90	-	-	-	-	-	-	-	-	-	-	2
		She	27N	119E	3002	10	3	3	4	0.45	0.55	-	-	-	-	-	-	-	-	-	-	2
		Tu	36N	101E	4003	10	0	2	8	0.10	0.90	-	-	-	-	-	-	-	-	-	-	2
		Tujia	29N	109E	3225	10	2	3	5	0.35	0.65	-	-	-	-	-	-	-	-	-	-	2
		Uygur	44N	81E	4893	10	0	1	9	0.05	0.95	-	-	-	-	-	-	-	-	-	-	2
		Xibo	43-44N	81-82E	4837	9	0	4	5	0.22	0.78	-	-	-	-	-	-	-	-	-	-	2
		Yizu	28N	103E	3113	10	1	2	7	0.20	0.80	-	-	-	-	-	-	-	-	-	-	2
	Japan	Japanese	-	-	-	200	14	65	121	0.23	0.77	200	0	0	1.00	0.00	-	-	-	-	-	8
		Japanese	38N	138E	4225	31	2	10	19	0.23	0.77	-	-	-	-	-	-	-	-	-	-	2
		Japanese in Tokyo	-	-	-	113	5	48	60	0.26	0.74	112	1	0	0.996	0.004	-	-	-	-	-	3
	South Korea	Koreans	-	-	-	162	-	-	-	0.19	0.81	-	-	-	1.00	0	-	-	-	0.97	0.03	7
	Malaysia	Malay	-	-	-	98	10	56	32	0.39	0.61	98	0	0	1.00	0.00	-	-	-	-	-	5
South Asia	India	Indians	-	-	-	90	11	51	28	0.41	0.59	90	0	0	1.00	0.00	-	-	-	-	-	5
	Pakistan	Balochi	66-67N	30.31E	3370	25	0	10	15	0.20	0.80	-	-	-	-	-	-	-	-	-	-	2
		Brahui	66-67N	30.31E	3370	25	0	6	19	0.12	0.88	-	-	-	-	-	-	-	-	-	-	2
		Burusho	36-37N	73-75E	4059	25	3	5	17	0.22	0.78	-	-	-	-	-	-	-	-	-	-	2
		Hazara	33-34N	70E	3725	24	0	12	12	0.25	0.75	-	-	-	-	-	-	-	-	-	-	2
		Kalash	35-37N	71-71E	4003	25	1	10	14	0.24	0.76	-	-	-	-	-	-	-	-	-	-	2
		Makrani	26N	66-66E	2891	25	0	7	18	0.14	0.86	-	-	-	-	-	-	-	-	-	-	2
		Pathan	32-35N	69-72E	3725	25	0	6	19	0.12	0.88	-	-	-	-	-	-	-	-	-	-	2

		Sindhi	24-27N	68-70E	2835	25	1	7	17	0.18	0.82	-	-	-	-	-	-	-	-	-	-	2
Central America	Mexico	Maya	19N	91W	2113	24	4	6	14	0.29	0.71	-	-	-	-	-	-	-	-	-	-	2
		Pima	29N	108W	3225	25	5	17	3	0.54	0.46	-	-	-	-	-	-	-	-	-	-	2
	El Salvador and Nicaragua	Mestizo	-	-	-	232	7	96	129	0.24	0.76	-	-	-	-	-	-	-	-	-	-	15
South America	Brazil	Karitiana	10S	63W	1112	24	1	9	14	0.23	0.77	-	-	-	-	-	-	-	-	-	-	2
		Surui	11S	62W	1223	21	1	5	15	0.17	0.83	-	-	-	-	-	-	-	-	-	-	2
	Colombia	Colombians	3N	68W	333.6	13	0	4	9	0.15	0.85	-	-	-	-	-	-	-	-	-	-	2
	Colombia	Colombians from Medellian	-	-	-	60	16	25	19	0.48	0.52	59	1	0	0.98	0.02	-	-	-	-	-	18
	Ecuador	Mestizo	-	-	-	317	7	64	246	0.12	0.88	-	-	-	-	-	-	-	-	-	-	16
Europe	Numerous	European	-	-	-	22	0	1	21	0.02	0.98	0	0	21	1.00	0.00	22	0	0	1.00	0	4
	Armenia	Southern Armenians	40	45	5004	100	0	10	90	0.05	0.95	100	0	0	1.00	0.00	100	0	0	1.00	0.00	1
	France	French	46N	2E	5115	29	1	3	25	0.09	0.91	-	-	-	-	-	-	-	-	-	-	2
		French Basque	43N	0E	4781	24	0	2	22	0.04	0.96	-	-	-	-	-	-	-	-	-	-	2
		French Caucasians	-	-	-	51	0	8	43	0.08	0.92	51	0	0	1.00	0	51	0	0	1.00	0	13
	Finland	Finnish	-	-	-	93	19	45	29	0.45	0.55	100	0	0	1.00	0	-	-	-	-	-	18
	Numerous	Sephardi Jews	-	-	-	66	0	14	52	0.11	0.89	68	0	0	1.00	0.00	68	0	0	1.00	0.00	1
	Germany	German Caucasians	-	-	-	1210	11	143	1056	0.07	0.93	-	-	-	-	-	-	-	-	-	-	6
	Israel	Druze	32N	35E	3558	48	1	6	41	0.08	0.92	-	-	-	-	-	-	-	-	-	-	2
		Palestinian	32N	35E	3558	51	2	14	35	0.18	0.82	-	-	-	-	-	-	-	-	-	-	2
		Bedouin	31N	35E	3447	48	2	12	34	0.17	0.83	-	-	-	-	-	-	-	-	-	-	2
	Italy	Bergamo (Northern Italian)	46N	10E	5115	14	1	3	10	0.18	0.82	-	-	-	-	-	-	-	-	-	-	2
		Sardinia	40N	9E	4448	28	0	3	25	0.05	0.95	-	-	-	-	-	-	-	-	-	-	2
		Tuscan	43N	11E	4781	8	0	1	7	0.06	0.94	-	-	-	-	-	-	-	-	-	-	2
		Tuscans	-	-	-	102	0	11	91	0.05	0.95	101	1	0	0.995	0.005	-	-		-	-	3
	Orkney Islands	Orcadian	59N	3W	6561	16	1	3	12	0.16	0.84	-	-	-	-	-	-	-	-	-	-	2
	Russia	Russian	61N	39-41E	6783	25	0	4	21	0.08	0.92	-	-	-	-	-	-	-	-	-	-	2
		Adygei	44N	39E	4893	17	0	4	13	0.12	0.88	-	-	-	-	-	-	-	-	-	-	2
	Siberia	Yakut	62-64N	129-139E	7005	25	0	5	20	0.10	0.90	-	-	-	-	-	-	-	-	-	-	2
	Spain	Northern Spaniard	-	-	-	204	3	31	171	0.09	0.91	-	-	-	-	-	-	-	-	-	-	10
		Southern Spaniard	-	-	-	177	5	22	150	0.09	0.91	-	-	-	-	-	-	-	-	-	-	15
		Iberians	-	-	-	14	2	7	5	0.39	0.61	14	0	0	1.00	0	-	-	-	-	-	18
	Sweden	Swedes	-	-	-	136	1	16	119	0.07	0.93	136	0	0	1.00	0	136	0	0	1.00	0	7
	Turkey	Anatolians	35.00	39.00	4337	74	2	10	62	0.09	0.91	74	0	0	1.00	0	74	0	0	1.00	0.00	1
	United Kingdom	British in England and Scotland	-	-	-	89	13	37	39	0.35	0.65	89	0	0	1.00	0	-	-	-	-	-	18
	Numerous	Ashkenazi Jews	-	-	-	90	0	6	84	0.03	0.97	90	0	0	1.00	0.00	91	0	0	1.00	0.00	1
Oceania	Bougainville	NAN Melanesian	6S	155E	667.2	22	1	6	15	0.18	0.82	-	-	-	-	-	-	-	-	-	-	2
	New Guinea	Papuan	4S	143E	444.8	17	1	5	11	0.21	0.79	-	-	-	-	-	-	-	-	-	-	2
North America	Canada	Canadian Caucasians	-	-	-	77	-	-	-	0.07	0.93	-	-	-	1.00	0	-	-	-	1.00	0	14
	Puerto Rico	Puerto Ricans	-	-	-	55	16	30	9	0.56	0.44	49	6	0	0.95	0.05	-	-	-	-	-	18
	U.S.A.	African Americans	-	-	-	15	6	7	2	0.63	0.37	0	3	10	0.88	0.12	9	4	1	0.79	0.21	4
		African ancestry in S.W. USA	-	-	-	87	39	41	7	0.68	0.32	71	15	0	0.91	0.09	-	-	-	-	-	3
		Chinese in Denver, Colorado	-	-	-	109	6	42	61	0.25	0.75	-	-	-	-	-	-	-	-	-	-	3
		Gujarati Indians in Houston, Texas	-	-	-	101	6	38	57	0.25	0.75	-	-	-	-	-	-	-	-	-	-	3
		Hispanic	-	-	-	22	0	11	11	0.25	0.75	0	0	21	1.00	0.00	21	0	0	1.00	0	4

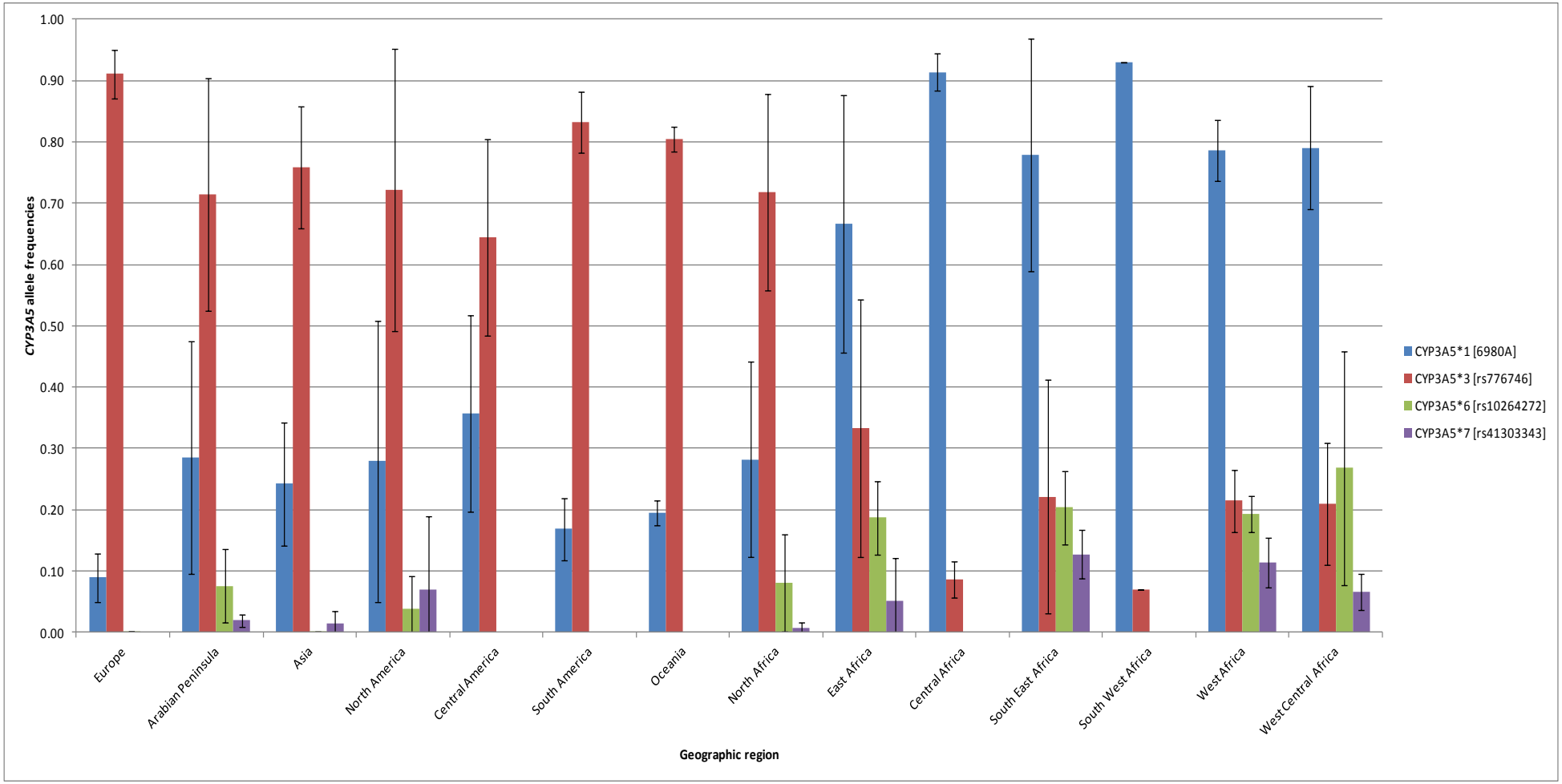
		Mexican ancestry in Los Angeles	-	-	-	85	5	32	48	0.25	0.75	82	4	0	0.98	0.02	-	-	-	-	-	3
		North American Caucasians	-	-	-	437	3	74	358	0.09	0.90	-	-	-	-	-	-	-	-	-	-	12
		Northern and Western European ancestry (CEPH collection)	-	-	-	143	0	11	132	0.04	0.96	67	0	0	1.00	0	-	-	-	-	-	3
North Africa	Algeria	Algerians	-1.045	35.505	3948	161	9	42	108	0.19	0.81	146	15	0	0.95	0.05	159	2	0	0.99	0.01	1
		Mzab (Mozabite)	32N	3E	3558	29	0	9	20	0.16	0.84	-	-	-	-	-	-	-	-	-	-	2
	Morocco	Berber	-6.84	34.03	3784	85	3	28	54	0.20	0.80	79	7	0	0.96	0.04	85	1	0	0.99	0.01	1
	Sudan	Kordofan	30.35	13.08	1454	30	11	11	8	0.55	0.45	19	10	1	0.80	0.20	29	1	0	0.98	0.02	1
		Northern	32.53	15.59	1734	133	24	58	51	0.40	0.60	104	28	0	0.89	0.11	135	1	0	1.00	0.00	1
	Tunisia	Tunisian	-	-	-	36	0	14	22	0.19	0.81	35	1	0	0.99	0.01	36	0	0	1.00	0	13
Central Africa	Central African Republic	Biaka Pygmies	4N	17E	444.8	33	28	3	2	0.89	0.11	-	-	-	-	-	-	-	-	-	-	2
	D.R. Congo	Mbuti Pygmies	1N	29E	111.2	15	13	2	0	0.93	0.07	-	-	-	-	-	-	-	-	-	-	2
East Africa	Ethiopia	Afar	41.36039	11.60212	1290	73	10	31	32	0.35	0.65	47	26	0	0.82	0.18	73	0	0	1.00	0.00	1
		Amhara	38.65951	9.869192	1097	76	14	22	40	0.33	0.67	55	19	2	0.85	0.15	76	0	0	1.00	0.00	1
		Anuak	34.41219	7.953241	884.4	76	38	32	6	0.71	0.29	44	25	7	0.74	0.26	75	1	0	0.99	0.01	1
		Ethiopians	-	-	-	150	-	-	-	0.79	0.21	-	-	-	0.88	0.12	-	-	-	1.00	0	9
		Maale	36.64333	5.714812	635.5	75	20	36	19	0.51	0.49	53	22	0	0.85	0.15	74	1	0	0.99	0.01	1
		Oromo	37.30817	7.83736	871.5	74	12	28	34	0.35	0.65	55	19	1	0.86	0.14	75	0	0	1.00	0.00	1
	Kenya	Bantu NE	3S	37E	333.6	12	9	2	1	0.83	0.17	-	-	-	-	-	-	-	-	-	-	2
		Luhya (Webuye, Kenya)	-	-	-	110	83	24	3	0.86	0.14	61	39	9	0.74	0.26	-	-	-	-	-	3
		Maasai (Kinyawa, Kenya)	-	-	-	184	45	97	42	0.51	0.49	137	43	4	0.86	0.14	-	-	-	-	-	3
	Sudan	Southern	31.77	5.18	576	125	74	42	9	0.76	0.24	58	50	15	0.67	0.33	117	8	0	0.97	0.03	1
	Tanzania	Chagga	38.05	-5.38	598.2	50	28	18	4	0.74	0.26	36	14	0	0.86	0.14	41	9	0	0.91	0.09	1
		Tanzanians	-	-	-	143	90	52	1	0.81	0.19	95	42	6	0.81	0.19	111	31	1	0.88	0.12	7
	Uganda	Bantu speakers	31.45	-0.57	63.38	39	36	3	0	0.96	0.04	22	17	0	0.78	0.22	23	16	0	0.79	0.21	1
		Ugandans	-	-	-	140	-	-	-	0.82	0.18	-	-	-	0.83	0.17	-	-	-	0.89	0.11	11
South East Africa	Malawi	Chewa	33.78	-13.98	1555	92	66	25	1	0.85	0.15	66	23	3	0.84	0.16	60	31	0	0.83	0.17	1
		Lomwe	33.78	-13.98	1555	18	13	4	1	0.83	0.17	10	8	0	0.78	0.22	14	4	0	0.89	0.11	1
		Other Malawians	33.78	-13.98	1555	14	9	4	1	0.79	0.21	10	4	0	0.86	0.14	10	4	0	0.86	0.14	1
		Ngoni	33.78	-13.98	1555	18	15	2	1	0.89	0.11	9	6	3	0.67	0.33	16	2	0	0.94	0.06	1
		Tumbuka	33.92	-11.45	1273	62	44	18	0	0.85	0.15	40	17	5	0.78	0.22	45	17	0	0.86	0.14	1
		Yao	33.78	-13.98	1555	56	37	18	1	0.82	0.18	43	12	1	0.88	0.13	46	10	0	0.91	0.09	1
	Mozambique	Sena	35.05	-17.44	1939	82	58	21	3	0.84	0.16	51	28	5	0.77	0.23	59	25	1	0.84	0.16	1
	South Africa	Bantu speakers	28.23	-25.71	2859	49	27	20	2	0.76	0.24	29	9	3	0.82	0.18	34	4	2	0.90	0.10	1
	Zimbabwe	Lemba	29.075	-23.095	2568	23	17	6	0	0.87	0.13	13	10	1	0.75	0.25	17	7	0	0.85	0.15	1
		Shona	-	-	-	100	-	-	-	0.22	0.78	-	-	-	0.78	0.22	-	-	-	0.90	0.10	14
		Zimbabweans	30	-19.67	2187	47	36	7	4	0.84	0.16	36	10	3	0.84	0.16	34	16	2	0.81	0.19	1
South West Africa	Namibia	San	21S	20E	2335	7	6	1	0	0.93	0.07	-	-	-	-	-	-	-	-	-	-	2
West Africa	Gabon	Gabonese	-	-	-	36	21	15	0	0.79	0.21	14	7	0	0.81	0.19	14	7	0	0.81	0.19	13
	Gambia	Gambians	-	-	-	288	183	90	15	0.79	0.21	179	100	9	0.80	0.20	222	62	4	0.88	0.12	17
	Ghana	Asante	6.17	-0.55	686.1	35	27	8	0	0.89	0.11	20	13	1	0.78	0.22	29	5	0	0.93	0.07	1
		Bulsa	10.73	-1.29	1193	90	58	29	3	0.81	0.19	61	28	0	0.84	0.16	69	19	2	0.87	0.13	1
		Kasena	10.89	-1.09	1211	47	28	17	2	0.78	0.22	31	16	0	0.83	0.17	35	12	0	0.87	0.13	1
	Senegal	Mandenka	12N	12W	1334	24	12	9	3	0.69	0.31	-	-	-	-	-	-	-	-	-	-	2
		Manjak	-15.88	12.986	1444	90	57	29	4	0.79	0.21	59	24	9	0.77	0.23	81	13	0	0.93	0.07	1
		Wolof	-17.453	14.687	1633	94	55	31	8	0.75	0.25	58	31	1	0.82	0.18	78	15	1	0.91	0.09	1
West Central	Cameroon	Kotoko	14.5	13	1446	39	18	21	0	0.73	0.27	23	16	1	0.78	0.23	36	4	0	0.95	0.05	1

Africa																					
		Lake Chad other	14.5	13	1446	23	12	11	0	0.76	0.24	9	12	1	0.68	0.32	20	3	0	0.93	0.07	1
		Mayo Darle	11.55	6.47	719.4	117	66	38	13	0.73	0.27	71	33	13	0.75	0.25	102	15	0	0.94	0.06	1
		Shewa Arab	14.5	13	1446	69	26	31	12	0.60	0.40	42	24	3	0.78	0.22	60	9	0	0.93	0.07	1
		Somie	12.5	6	667.2	65	36	28	1	0.77	0.23	44	19	2	0.82	0.18	52	13	0	0.90	0.10	1
	Congo	Congolese	15.28	-4.26	473.7	55	35	18	2	0.80	0.20	43	11	1	0.88	0.12	45	10	0	0.91	0.09	1
	Nigeria	Igbo	8.32	4.95	550.4	87	64	23	0	0.87	0.13	60	24	4	0.82	0.18	73	12	2	0.91	0.09	1
		Yoruba	-	-	-	198	130	60	3	0.83	0.17	134	53	6	0.83	0.17	-	-	-	-	-	3
		Yoruba	6-10N	2-8E	889.6	25	9	3	0	0.94	0.06	-	-	-	-	-	-	-	-	1.00	0	2
		Yoruba	-	-	-	12	22	3	0	0.88	0.13	1	2	5	0.25	0.75	11	0	0	-	-	4

List of references for Table 3.10:

1. R Bains, unpublished data from this study,
2. (Thompson et al. 2004),
3. The International HapMap project (http://hapmap.ncbi.nlm.nih.gov/),
4. NIEHS SNPs program (http://egp.gs.washington.edu/),
5. (Balram et al. 2003),
6. (Dally et al. 2004),
7. (Diczfalusy et al. 2008),
8. (Fukuen et al. 2002),
9. (Gebeyehu et al.),
10. (Floyd et al. 2003),
11. (Mukonzo et al.),
12. (Plummer et al. 2003),
13. (Quaranta et al. 2006),
14. (Roy et al. 2005),
15. (Sinues et al. 2007),
16. (Sinues et al. 2008),
17. (Wojnowski et al. 2004)
18. 1000 Genomes genotyping data, from the Illumina Omni platform, for populations that are not part of the HapMap database.

Figure 3.7: A graph showing the average *CYP3A5* allele frequencies by geographic region. The Figure shows the average frequencies for each geographic region listed in Table 3.8. Error bars denote standard deviation.



3.3.5 The association between CYP3A5 allele frequencies and latitude

A previous study reported that *CYP3A5*3* allele frequencies are positively correlated with increased distance from the equator (Thompson et al. 2004). The genotype data for the *CYP3A5*1/CYP3A5*3* locus for the 36 populations genotyped in this study were combined with data for 1028 individuals from 51 populations genotyped in the original paper (*n*=3570 from 87 populations) to test for a global correlation between *CYP3A5*1* and *CYP3A5*3* allele frequencies and latitude (Figure 3.8 and 3.9).

A significant global positive correlation between increased distance from the equator and *CYP3A5*3* allele frequencies was observed when all 87 populations were examined [Spearmann Rho = 0.701, $p<0.0001$]. Analysis of the seven geographic regions represented by the 87 populations, where more than ten population groups had been sampled, found a significant positive correlation between latitude and *CYP3A5*3* allele frequencies in Asia [Spearmann Rho = 0.853, $p<0.001$] but not within Europe [Spearmann Rho = 0.002, $p=0.994$] or Africa [Spearmann Rho = 0.277, $p=0.08$]. Although removal of the heterogeneous Ethiopian populations from the analysis found a significant positive correlation in Africa [Spearmann Rho = 0.369, $p=0.04$]. The lack of an association between latitude and allele frequencies in Europe is due to homogeneity in allele frequencies between population groups from the region.

There is a strong positive correlation between north latitude and *CYP3A5*3* allele frequencies [Spearmann Rho = 0.781, $p<0.0001$] but not for south latitude [Spearmann Rho = -0.318, $p=0.199$]. Due to the overrepresentation of north latitude populations in both datasets (69 populations out of a total of 87), the analysis was performed using only data obtained for African groups. The results confirmed those for 87 populations; *CYP3A5*3* frequencies are positively correlated with north latitude (26/40 African groups) [Spearmann Rho = 0.621, $p<0.0001$], but not with south latitude (14/40 African groups) [Spearmann Rho = -0.07, $p=0.820$].

Figure 3.8: A map showing the distribution of the *CYP3A5*1/CYP3A5*3* alleles in each of 87 populations; 51 of which were previously genotyped (Thompson et al. 2004), and 36 genotyped for this study. The size of each circle is proportional to the number of individuals sampled from a given population.

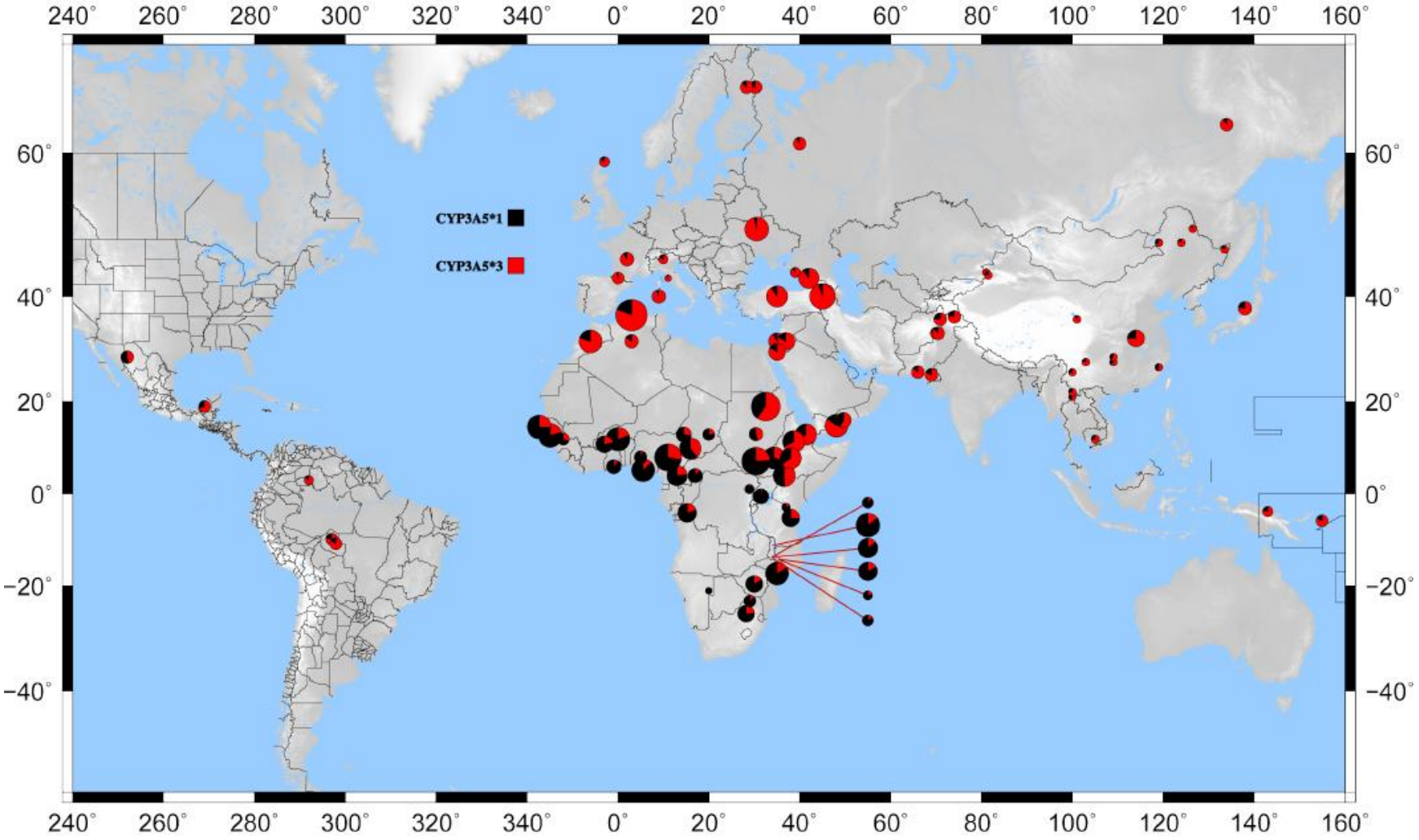
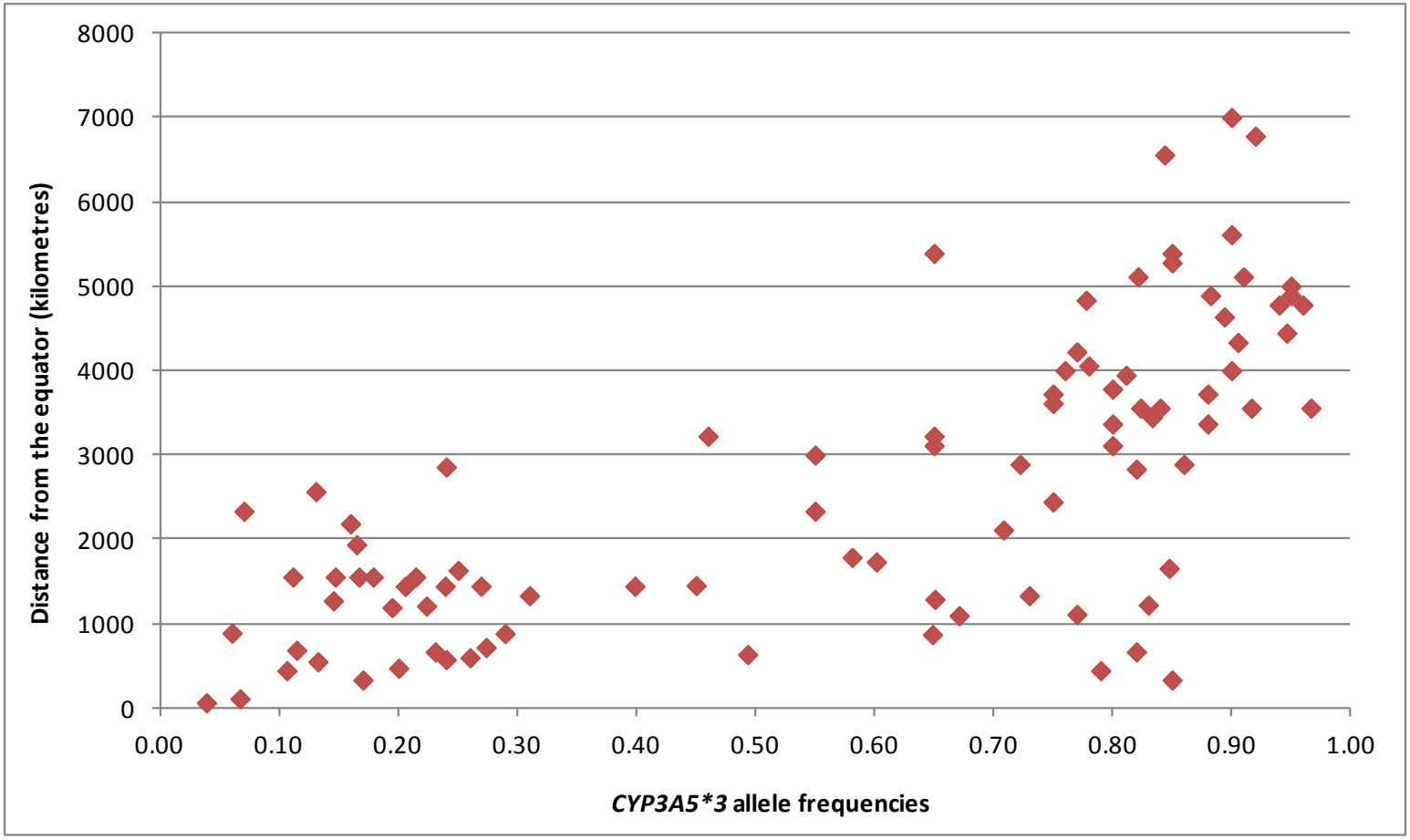


Figure 3.9: A scatter plot showing the correlation between *CYP3A5*3* allele frequencies and distance from the equator in kilometres**.** The allele frequencies are for the combined dataset of 3570 individuals from 87 global populations.



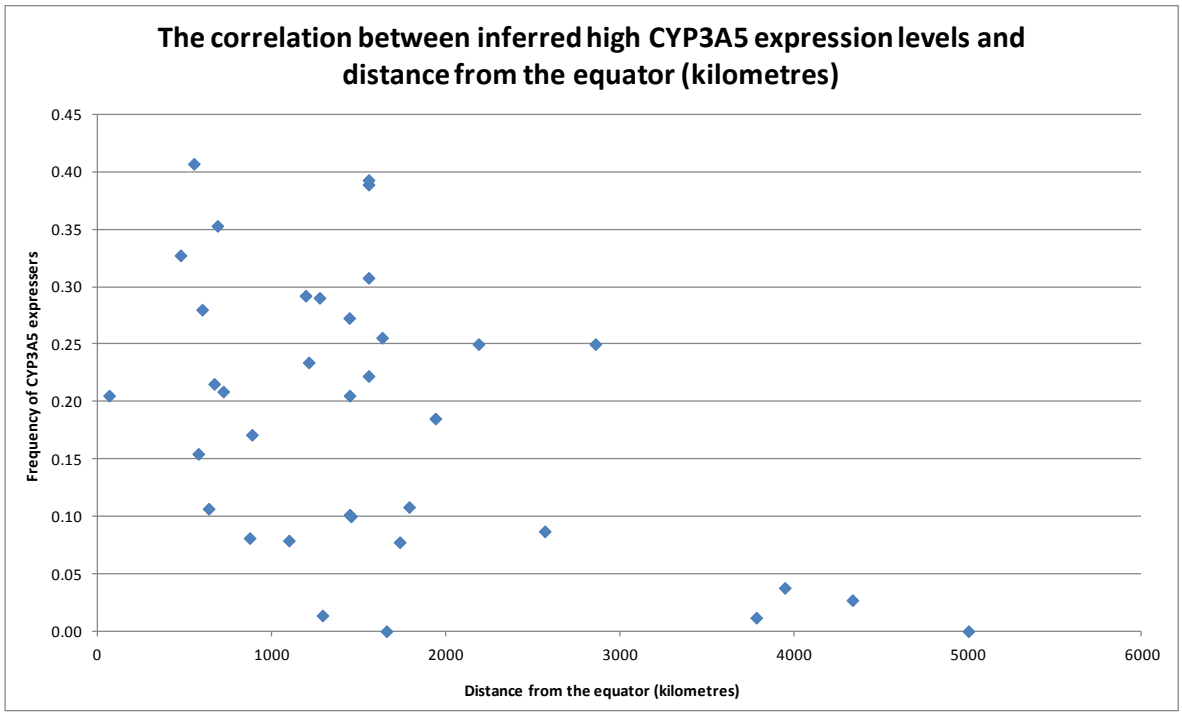
There is a strong global trend towards *CYP3A5*3* frequencies increasing with increased distance from the equator.

An examination of correlations between *CYP3A5*6* and *CYP3A5*7* alleles was performed for the 36 populations genotyped for this thesis. A significant negative correlation between *CYP3A5*6* frequencies and geographic distance from the equator was observed [Spearmann Rho = -0.35, p=0.03]; although this correlation was no longer significant when only groups south of the equator were examined [Spearmann Rho = 0.237, p=0.163]. No significant correlation between *CYP3A5*7* frequencies and distance from the equator was observed [Spearmann Rho = -0.089, p=0.603]. This is almost certainly influenced by the confined distribution of the *CYP3A5*7* allele to Niger-Congo speaking groups; who are homogeneous and spread over large geographic distances.

3.3.6 The correlation between CYP3A5 expression phenotype and latitude

A Spearmann's Rank correlation test was performed to examine whether there is an association between CYP3A5 expression phenotypes inferred from diplotypes (Figure 3.10). Significantly negative correlations between inferred high- and intermediate- CYP3A5 expression phenotypes and distance from the equator were observed from the data [Spearmann Rho = -0.404, p<0.02] and [Spearmann Rho = -0.44, p=0.007] respectively. In contrast to low/non- protein expression [Spearmann Rho = 0.454, p<0.006]. There was a strong negative correlation between the frequencies of expresser haplotypes and distance from the equator [Spearmann Rho = -0.490, p=0.002] and a strong positive correlation between the collective frequencies of low/non-expresser haplotypes and distance from the equator [Spearmann Rho = 0.490, p=0.002] (Figure 3.11). The results suggest that latitude is a predictor of CYP3A5 expression profiles.

Figure 3.10: A scatter plot showing the correlation between geographic distance from the equator in kilometres and inferred CYP3A5 expression phenotypes.



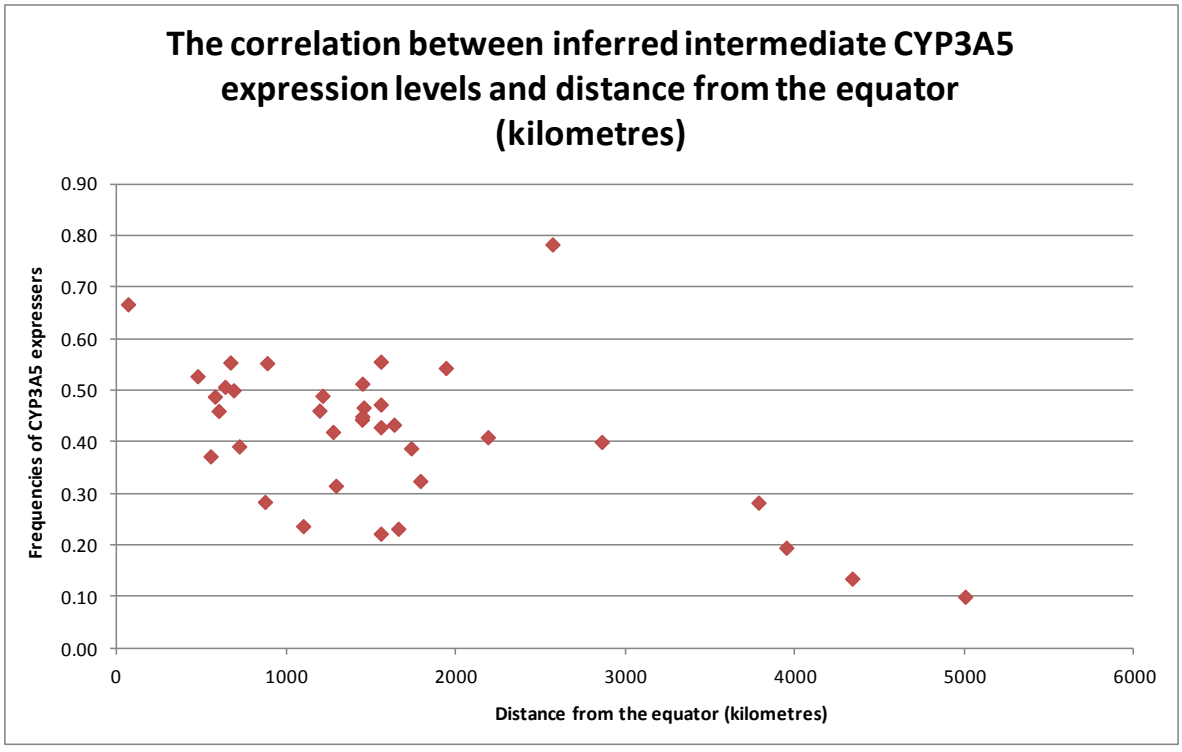


Figure 3.10 continued; the correlation between geographic distance from the equator (in kilometres) and inferred CYP3A5 expression phenotypes.

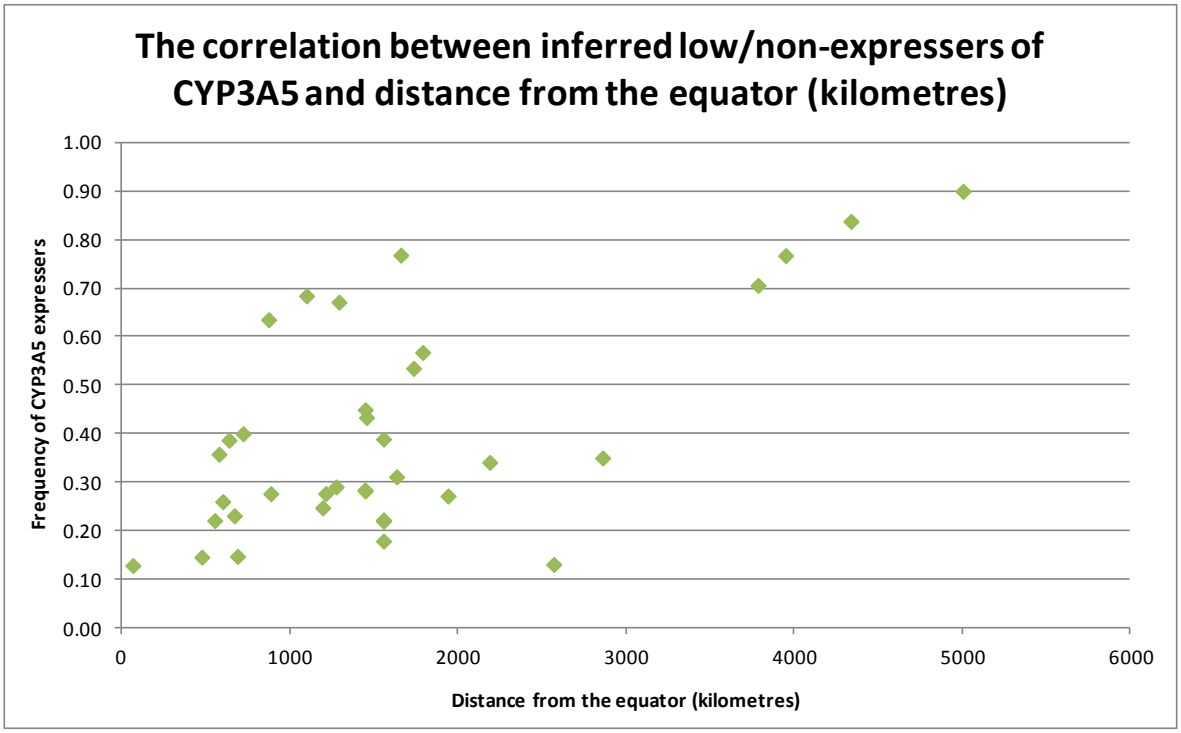
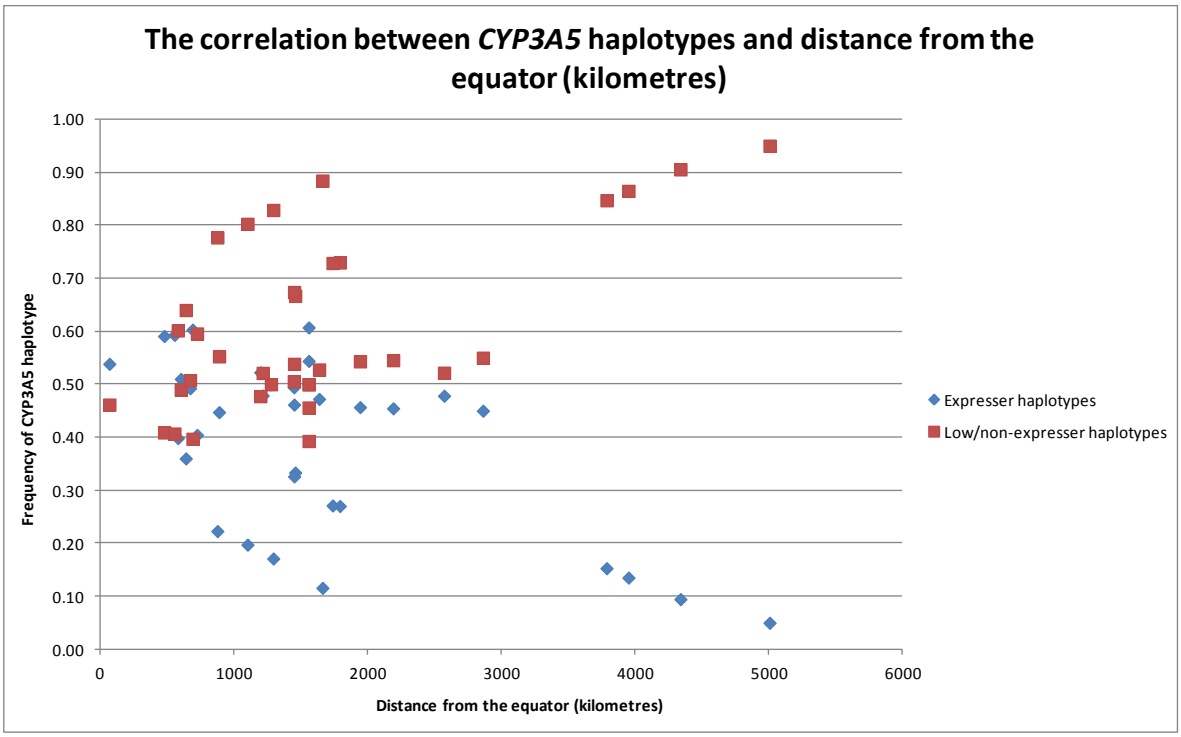


Figure 3.11: The correlation between CYP3A5 haplotypes (expresser and collective low/non-expresser) and distance from the equator (kilometres).



3.4 Discussion

3.4.1 Intra-African CYP3A5 expression levels are likely to be highly variable

In this chapter the distribution of *CYP3A5* variants, previously reported to affect protein expression, were genotyped in African populations. The results from this study are consistent with previous reports that the highest frequencies of *CYP3A5*1* are found in populations with recent African ancestry (Roy et al. 2005; Quaranta et al. 2006). The results also indicate that the *CYP3A5*6* and *CYP3A5*7* variants are found at high frequencies across Africa, suggesting that multiple variants are likely to affect CYP3A5 expression levels across the continent. Interestingly the three low/non expresser variants occur on independent haplotype backgrounds suggesting that traits causing low/non CYP3A5 expression in Africa have evolved more than once.

Previous studies have estimated that between 55-95% of individuals with recent African ancestry have high levels of CYP3A5 expression (Hustert et al. 2001; Kuehl et al. 2001; Lamba et al. 2002). One explanation for the large reported range of CYP3A5 expressers is the choice of *CYP3A5* markers used to infer protein expression in previous studies. The majority of previous studies have focused on the *CYP3A5*1/CYP3A5*3* locus as a single determinant of CYP3A5 expression. Genotyping of the *CYP3A5*1/CYP3A5*3* locus alone may be sufficient to predict CYP3A5 expression profiles in non African-American and African populations. However the high frequencies of *CYP3A5*6* and *CYP3A5*7* in Africa, reported in this chapter, reiterate the importance of genotyping these variants in populations with recent African ancestry.

The classification of *CYP3A5* alleles as expresser or non-expresser is likely to have an effect on the range of reported expresser frequencies in Africa. Of the three low/non-expresser variants genotyped in this study, there is some doubt over the functional implications of the *CYP3A5*6* variant. The effect of *CYP3A5*6* on exon 7 skipping (see section 1.2.3.1) has only been established once in 2001 (Kuehl et al. 2001) and has not been independently replicated. Although the authors reported that *CYP3A5*6* caused skipping of exon 7, they observed that normally spliced mature mRNA was present in all individuals who carried the *CYP3A5*6* variant (Kuehl et al. 2001). The study only included three individuals who were heterozygous for the *CYP3A5*6* mutation and therefore it is not known whether normally spliced mature mRNA is observed in individuals who are homozygous for *CYP3A5*6*; and if so, at what concentration comparative to *CYP3A5*1* homozygotes.

One method by which the functionality of *CYP3A5*6* could be established is to examine clinical data on a drug substrate of CYP3A5, such as the well studied immunosuppressant tacrolimus. The effect of low/non-CYP3A5 expression, determined by genotypes at the *CYP3A5*1/CYP3A5*3* locus, on tacrolimus associated adverse drug reactions is well established (Goto et al. 2004; Quteineh et al. 2008). Therefore patients, particularly those with recent African ancestry, who are not heterozygous or homozygous for *CYP3A5*3* and yet display the same adverse clinical outcomes as *CYP3A5*3* carriers should be genotyped for the *CYP3A5*6* variant. If a significant number of patients who are *CYP3A5*6* carriers have adverse clinical outcomes then it would provide evidence of an association between *CYP3A5*6* and adverse clinical outcomes that are known to be associated with low/non CYP3A5 expression.

To establish the effect of *CYP3A5*6* on mRNA splicing, an *in vitro* splicing assay in which double stranded DNA known to contain the *CYP3A5*6* variant could be cloned into a bacterial vector, transformed into bacterial culture and transfected into eukaryotic cells using established techniques (Webb et al. 2003). Following incubation and extraction of mRNA for analysis by reverse transcription, an examination and comparison of the cDNA sequence from the *CYP3A5*6* clones alongside a control, from a homozygous non-*CYP3A5*6* clone would be performed. If *CYP3A5*6* causes exon skipping then the cDNA sequence, obtained from *CYP3A5*6* clones, would not contain the sequence for exon 7 of *CYP3A5*.

The results from either method could change the classification of *CYP3A5*6,* as a low/non-expresser variant, to a variant which can reduce protein expression or a variant which has no effect on protein expression and processing at all. This could alter which *CYP3A5* variants are considered to be clinically relevant.

Another reason for the wide range of reported CYP3A5 expresser frequencies are that African populations have been underrepresented in previous studies on CYP3A5 variability. A large number of studies have estimated CYP3A5 expression levels in sub-Saharan Africa by extrapolating data obtained for African-American populations. The proportion of individuals expressing CYP3A5 protein, at high concentrations, is expected to be lower in African-American populations comparative to sub-Saharan Africans. The rationale of expecting higher frequencies of CYP3A5 expressers in sub-Saharan African populations is due to Caucasian admixture in African-American populations (Reed 1969; Destro-Bisol et al. 1999).

The data from this study are consistent with previous expectations that the frequencies of CYP3A5 expressers in Africa are higher than in other global populations. However, the proportion of Africans who express CYP3A5, at high concentrations, is likely to be much lower than the previous estimate of 95%. Additionally expression phenotypes are likely to be highly variable across Africa; sub-Saharan Africans have higher frequencies of

CYP3A5 expressers than North Africans. There are also considerable differences between East Africans and other sub-Saharan populations. The East African populations typed in this study have substantially lower levels of predicted CYP3A5 expression than in other regions of sub-Saharan Africa, additionally there is substantial heterogeneity within Ethiopia, consistent with a previous study of *CYP3A5* variability within the country (Gebeyehu et al. 2011). This demonstrates that the method of inferring African CYP3A5 expression phenotypes from African-American data does not account for the considerable heterogeneity observed across the continent.

The data strongly suggest that language family and north latitude are stronger predictors of CYP3A5 expression phenotype than African ancestry or geographic region alone. Individuals predicted to express CYP3A5 at high concentrations are overrepresented in equatorial and Niger-Congo speaking populations. These findings are consistent with previous reports that there is a correlation between African genetic diversity and language (Wood et al. 2005; Reed and Tishkoff 2006). Within language families, there was less diversity in *CYP3A5* allele frequencies between Niger-Congo speaking populations than for other language groups in Africa. The similarities between Niger-Congo speaking populations are influenced by the recent expansion of Bantu-speaking populations ~4000 years ago (Beleza et al. 2005). This is consistent with previous studies of genetic diversity in Africa; non Niger-Congo speaking groups have greater intra-population diversity (Wood et al. 2005) and this is also true of diversity at the *CYP3A5* locus.

3.4.2 The potential implications of CYP3A5 variability in Africans

As outlined in section 1.1.4, sub-Saharan Africa has a high burden of diseases which are treated with a wide spectrum of drugs (Aspray et al. 1998; Coleman 1998). The results from this study show that there is variability in the expression of an important drug metabolising enzyme; responsible for the metabolism of a wide spectrum of clinical drugs (Lamba et al. 2002). CYP3A5 variability has been previously reported as having significant effects on treatment of many drugs (see section 1.2.4) including those used in the treatment of malaria, (Ferreira et al. 2008) HIV-1, (Josephson et al. 2007) cancer (Dandara et al. 2005) heart disease (Bochud et al. 2006) and immunosuppressants, (Quteineh et al. 2008); diseases which are common in populations with recent African ancestry.

What is clear from the data reported in this chapter is that the approach of grouping African populations in translational medical research would overlook, potentially fatally, the substantial diversity within the continent; particularly between East Africans and other sub-

Saharan Africans. Intra-African differences in the prevalence of clinically relevant alleles mean that standardising dosages of CYP3A5 substrates within these populations on the basis of skin colour would not benefit individuals in a way that methods accounting for population-specific variability, such as language family or geographic region of ancestry, would.

3.4.3 Natural selection at the CYP3A5 locus

One of the most striking features of the results is the large disparity in *CYP3A5*1/CYP3A5*3* frequencies between African and non-African populations. Previous studies have reported that the Sahara desert acts as a major barrier to gene flow between sub-Saharan Africans and other global populations (Cruciani et al. 2002; Salas et al. 2002). The substantial differences in *CYP3A5* allele frequencies between North and sub-Saharan Africans (reported in this study) are consistent with, but not entirely explained by, the Sahara barrier.

The results from Africa fit a correlation from a previous report (Thompson et al. 2004) that *CYP3A5*3* frequencies are positively correlated with increased distance from the equator. However the correlation appears to be specific to increased north latitude. The ratio of north to south latitude groups in this study is ~4:1 which may have skewed the initial observation that latitude is correlated with *CYP3A5*1/CYP3A5*3* frequencies. However when only African groups were considered, the ratio of north to south latitude African groups is ~1:1, there was a correlation between *CYP3A5*1/CYP3A5*3* frequencies and north latitude.

One of the most interesting features of the data reported in this chapter and those previously published (Thompson et al. 2004) is that the distribution of *CYP3A5*1/CYP3A5*3* alleles are identical to functionally important variants of genes involved in blood pressure regulation (Young et al. 2005). The frequency of the non-expresser, and protective against elevated blood pressure, *CYP3A5*3* allele is positively correlated with distance from the equator. The ancestral *CYP3A5*1* allele has been reported to be associated with increased systolic blood pressure and mean arterial pressure in African-American populations (Givens et al. 2003). Hypertension is highly prevalent in Africa (Cooper et al. 1997; Sobngwi et al. 2002) and populations with recent African ancestry are overrepresented in hypertension patient populations. Although there are environmental and dietary factors which contribute to the risk of developing hypertension risk, genetic factors also play a major role in disease pathology (Young et al. 2005).

CYP3A5 has been proposed to catalyse the conversion of cortisol to 6-β-hydroxycortisol in the kidney, which leads to higher sodium reabsorption and water retention (Ghosh et al. 1995). This mechanism is vital for populations who experience frequent water

shortages; namely those closest to the equator and those who are prone to salt loss through excessive sweating. This potentially explains the high prevalence of *CYP3A5*1* in equatorial populations, previously studied, and seen in this study. Conversely, as the distance from the equator increases, temperature and humidity decrease and there is less selective pressure to retain water and reabsorb sodium.

Comparisons of hypertension patients with recent African ancestry and other global populations suggest that the *CYP3A5*1* allele may provide an advantage near the equator but is detrimental at north latitudes. In contrast, the *CYP3A5*3* allele appears to be protective against, genetically determined, elevated blood pressure, and this is perhaps why the allele has risen to such high frequency in these populations.

What is notable however is the *CYP3A5*3* allele has almost reached fixation in populations outside of Africa and considerably north of the equator. There may be evidence of selection for the low/non-functional *CYP3A5*3* allele outside of Africa. Evidence for this would come from an examination of the haplotype backgrounds on which the non-functional variant occurs. Any new mutation will occur on a specific haplotype background, under neutrality we would expect to see the mutation associated with a subset of all haplotypes and its frequency would vary according to genetic drift. However if the allele becomes advantageous and is selected for, it will rise in frequency and all tightly linked and associated variation, on a particular haplotype background, will be selected for. If this is fairly recent then there will be an excess of rare mutations on the haplotype containing the selected mutation and the haplotype will have far less diversity than expected under a neutral model of mutation, a so called mutation-selection-balance model (Di Rienzo and Hudson 2005). The following chapters will examine this in greater detail by examining the full *CYP3A5* gene sequence obtained for different global populations to examine evidence of selection in and outside of Africa.

As for the *CYP3A5*6* and *CYP3A5*7* variants, the geographic distribution of *CYP3A5*7* appears to be identical to a major demographic event that occurred within Africa approximately 4000 years ago; known as the Expansion of the Bantu Speaking Peoples (Tishkoff and Verrelli 2003; Beleza et al. 2005). *CYP3A5*7* was observed at high frequencies within West, West Central and South East Africa which would be consistent with the second wave of the Expansion of Bantu Speaking Peoples, and this suggests an origin of *CYP3A5*7* prior to this event ~4000 years ago. It is likely that *CYP3A5*7* evolved in West or West Central Africa and spread through Africa as a result of human migration patterns. What is interesting about this variant is that, whilst its distribution is restricted, it appears to have evolved in populations close to the equator and has a null effect on CYP3A5 expression. It is possible that

it may have evolved to help regulate the water retention and salt reabsorption effects of the *CYP3A5*1* allele; which may be detrimental if unregulated, even in equatorial populations.

In contrast the geographic distribution of *CYP3A5*6* is much more consistent across sub-Saharan Africa and suggests that this variant evolved before *CYP3A5*7* and had increased time to spread across the region. This variant is observed at low frequencies in the Yemeni and North African sample sets, although this may be as a result of the Arab slave trade during the 8-19th centuries (Richards et al. 2003). It is possible that *CYP3A5*6* and *CYP3A5*7* evolved to regulate too much sodium reabsorption and water retention in equatorial populations, this may explain their relatively low frequencies across the continent, in contrast to the high frequencies of *CYP3A5*3* observed outside of Africa.

3.4.4 Conclusions

Sub-Saharan Africans have been largely underrepresented in evolutionary and medical studies (Campbell and Tishkoff 2008). The data from this chapter aim to correct this by reporting data on functionally important variants of a medically important gene, *CYP3A5*. The significant positive correlation between latitude and decreased CYP3A5 expression is consistent with a hypothesis of positive selection for low/non expression of CYP3A5 in populations who reside at large distances north of the equator. Variability in CYP3A5 expression is likely to have important implications for the specificity of drugs and dosages used to treat patients with recent African ancestry, and in predicting their susceptibility to many diseases; such as hypertension. The following chapters aim to characterise intra-African *CYP3A5* variability further by identifying novel, and potentially functionally important, *CYP3A5* variants, and evaluate evidence of positive selection on the gene.

4. Intra-African diversity at the *CYP3A5* gene

4.1 Overview and specific aims of the chapter

In chapter 3, considerable intra-African differences in the frequencies of clinically relevant *CYP3A5* alleles were identified. Principal co-ordinates (PCO) analysis found that differences in *CYP3A5*3* allele frequencies account for over 95% of all inter-population variation. This chapter is a more restricted survey, than presented in chapter 3, and examines intra-African population structure in a 4448bp region of the *CYP3A5* gene.

The re-sequenced region includes exons 1, 2, 3, 4, 6, 7 and 11, plus the flanking introns. Thirteen diverse African populations were chosen for re-sequencing (see Table 4.1 and Figure 4.1). Details of primers, PCR and re-sequencing conditions are provided in chapter 2. All of the sample sets were genotyped as part of the geographic survey, presented in chapter 3. The populations represent four major global language families and a large geographic region within Africa. Given time and funding constraints it was not possible to re-sequence the entire *CYP3A5* gene in all thirteen African populations. However the regions encompassing the *CYP3A5*1/CYP3A5*3, CYP3A5*6* and *CYP3A5*7* loci were re-sequenced, as well as part of the proximal promoter in each population. Intra-African diversity was analysed in a global context with re-sequencing data for the same 4006 base pair region in three, ethnically diverse, North American populations from the Coriell repositories (see section 2.1.1.4.1). Ethiopian data for the entire *CYP3A5* gene are analysed in detail in chapter 5; in this chapter part of the dataset were included to analyse intra-African diversity.

4.2 *CYP3A5* variation

4.2.1 Intra-African diversity in the re-sequenced CYP3A5 region

20 polymorphic sites were identified within a 4448 base pair region (see Figure 4.2). Table 4.2 shows the full list of polymorphic sites identified in 8 non-Ethiopian African groups; Ethiopian re-sequencing data are presented in Table 5.6. Outside of Ethiopia all, except one 5´ UTR variant, occurred in the *CYP3A5* coding region. 18 variants were single nucleotide substitutions; one was a single base insertion in exon 11 which defines the *CYP3A5*7* variant and a novel 10 base pair deletion was observed in five heterozygous individuals. One variant was a singleton and only one, non-singleton, variant identified was specific to a population. Within an equivalent 4448 base pair region re-sequenced in five Ethiopian populations, a total of 26 variants were identified; 5 of which were singletons and 4 were specific to a population.

Table 4.1: A full list of sample sets in which *CYP3A5* was re-sequenced, including details on geographic location and language family. Some have been named according to ethnicity (*a*) and others according to where, geographically (*b*), the samples were collected.

Country	Sample set	Language family	Sample size	Latitude	Longitude	Distance from the equator (kilometres)	*CYP3A5*1* allele frequency	*CYP3A5*3* allele frequency	*CYP3A5*6* allele frequency	*CYP3A5*7* allele frequency
Ethiopia	Afar (*a*)	Afro-Asiatic	73	11.602	41.360	1290	0.35	0.65	0.18	0.00
	Amhara (*a*)	Afro-Asiatic	76	9.869	38.660	1097	0.33	0.67	0.15	0.00
	Maale (*a*)	Afro-Asiatic	75	5.715	36.643	635.5	0.51	0.49	0.15	0.01
	Oromo (*a*)	Afro-Asiatic	74	7.837	37.308	871.5	0.35	0.65	0.14	0.00
	Anuak (*a*)	Nilo-Saharan	76	7.953	34.412	884.4	0.71	0.29	0.26	0.01
Sudan	Kordofan (*b*)	Nilo-Saharan	30	13.08	30.35	1454	0.55	0.45	0.20	0.02
Ghana	Asante (*a*)	Niger-Congo A	34	5.82	-2.82	686.1	0.89	0.11	0.22	0.07
	Bulsa (*a*)	Niger-Congo A	22	10.73	-1.29	1193	0.81	0.19	0.16	0.13
Cameroon	Shewa Arabs (*a*)	Afro-Asiatic	65	15.05	12.11	1446	0.60	0.40	0.22	0.07
	Somie (*b*)	Niger-Congo A	65	6.00	12.5	667.2	0.77	0.23	0.18	0.10
Congo	Brazzaville (*b*)	Niger-Congo B	55	-4.26	15.28	473.7	0.80	0.20	0.12	0.09
Malawi	Chewa (*a*)	Niger- Congo B	50	-13.47	34.188	1555	0.85	0.15	0.16	0.17
Mozambique	Sena (*b*)	Niger-Congo B	51	-17.44	35.05	1939	0.84	0.16	0.23	0.16

Figure 4.1: A map showing the distribution of the thirteen sub-Saharan African populations re-sequenced in this study. The image has been adapted from taken from http://www.freeworldmaps.net/printable/africa/ using and edited by me using Microsoft PowerPoint 2007.

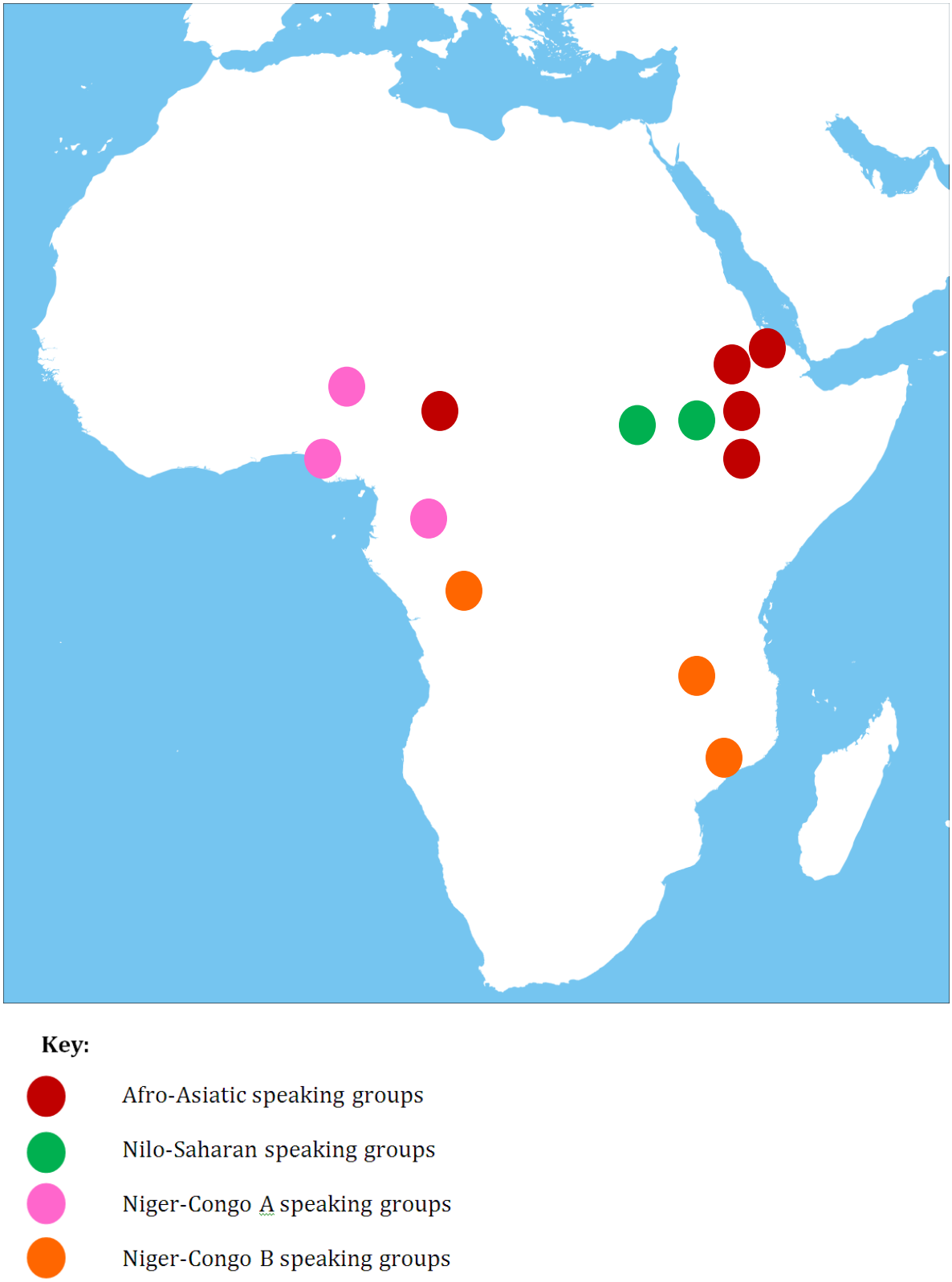


Figure 4.2: A full list of all polymorphic sites identified across the 4448 base pair region re-sequenced in eight, non-Ethiopian, African populations.

- The pink boxes represent re-sequenced *CYP3A5* exons, although they are not spaced according to scale. The red box represents the proximal promoter region of *CYP3A5.* Spacer regions, represented by black lines between adjacent exons, are introns. Grey boxes represent regions that were not re-sequenced in these populations.
- The exons are ordered from Exon 1-13, although they are not all numbered on the Figure.
- The numbers in the brackets correspond to the total number of variants observed across the 8 populations.

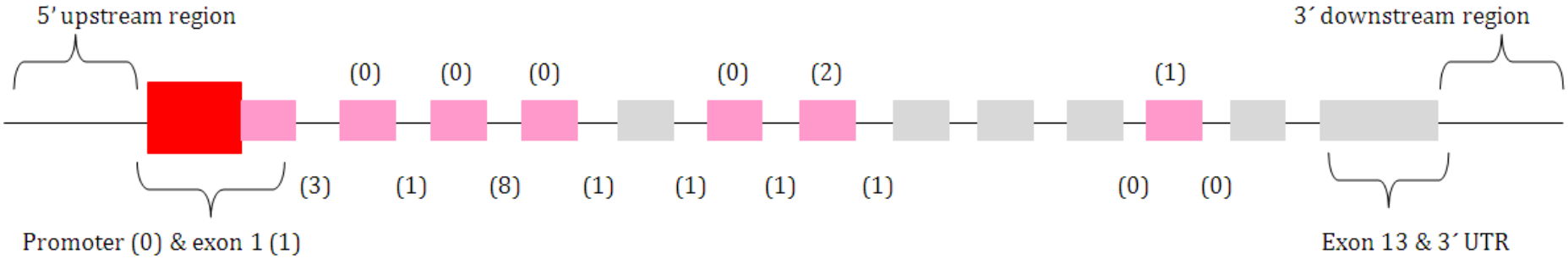


Table 4.2: A list of all polymorphic sites identified in a 4448 base pair region re-sequenced in 8 (non-Ethiopian) African populations. *f* is the frequency and *n* is the number of chromosomes and light blue shading indicates novel mutations

					Shewa Arabs		Congolese		Asante		Bulsa		Chewa		Sena		Mambila		Kordofan		
CYP3A5 region	Position on chromosome 7: (NCBI Build 132, February 2009)	*CYP3A5* variant and its position relative to the translation initiation codon (A of ATG is +1)	NCBI dbSNP database refSNP ID	Effect	*f*	n	*f*	n	*f*	n	*f*	n	*f*	n	*f*	n	*f*	n	*f*	n	Total re-sequenced
5´ UTR of exon 1	99277593	-74 C>T	rs28371764		0.00	1	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	3	666
Intron 1	99277445	74 10 base pair deletion			0.00	0	0.00	1	0.00	0	0.00	0	0.00	0	0.00	0	0.01	4	0.00	0	666
Intron 1	99277383	136 C>T			0.00	2	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	666
Intron 1	99277230	289 G>C			0.00	0	0.00	0	0.00	2	0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	666
Intron 2	99273701	3818 G>A			0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	668
Intron 3	99272310	5209 C>T	rs28365067		0.01	8	0.00	2	0.00	2	0.00	0	0.00	3	0.00	2	0.01	5	0.01	5	724
Intron 3	99272290	5229 G>A	rs41301652		0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00	1	0.00	3	0.00	0	724
Intron 3	99272009	5510 T>A	rs28969392		0.00	1	0.00	0	0.00	1	0.00	2	0.00	1	0.00	1	0.00	3	0.00	2	724
Intron 3	99271928	5591 C>T	rs41301655		0.00	1	0.00	3	0.01	6	0.01	5	0.00	2	0.00	1	0.01	10	0.00	0	724
Intron 3	99271853	5666 A>G	rs41301658		0.01	10	0.02	11	0.00	3	0.01	6	0.01	9	0.02	14	0.02	13	0.00	0	724
Intron 3	99271808	5711 A>G	rs41258334		0.01	7	0.00	2	0.00	1	0.00	0	0.00	3	0.00	1	0.01	4	0.01	5	724
Intron 3	99270539	6980 A>G	rs776746	Defines *CYP3A5*3*	0.07	53	0.03	22	0.01	8	0.01	7	0.02	13	0.02	16	0.04	30	0.04	27	744
Intron 3	99270318	7201 C>T	rs8175345		0.02	12	0.02	12	0.00	3	0.01	7	0.01	9	0.03	21	0.02	13	0.00	1	746
Intron 4	99270165	7354 C>T			0.00	0	0.00	0	0.00	0	0.00	0	0.01	4	0.00	0	0.00	0	0.00	2	746
Intron 5	99264391	13128 C>G			0.00	0	0.00	2	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.00	0	700
Intron 6	99264149	13370 G>A	rs41301670		0.00	3	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	700
Exon 7	99262835	14684 G>A	rs10264272	Defines *CYP3A5*6*	0.04	30	0.02	13	0.02	15	0.01	9	0.02	15	0.03	20	0.03	23	0.02	12	742
Exon 7	99262793	14726 A>G	rs28383472	Synonymous	0.00	0	0.00	2	0.01	5	0.01	4	0.00	2	0.00	0	0.01	8	0.00	0	742
Intron 7	99262689	14830 C>T			0.00	0	0.00	0	0.00	1	0.00	0	0.00	2	0.00	0	0.00	0	0.00	0	742
Exon 11	99250394	27125 1bp insertion	rs41303343	Defines *CYP3A5*7*	0.01	8	0.01	10	0.01	5	0.00	1	0.03	22	0.02	15	0.02	13	0.00	1	742

A cross-species alignment of the re-sequenced *CYP3A5* region with other primates, found that 16/20 polymorphic sites were in highly conserved nucleotide positions in primates. Highly conserved nucleotide positions are polymorphic sites where the ancestral allele, as inferred from the chimpanzee sequence, is identical in all primate species (Thompson et al. 2004; Thompson et al. 2006).

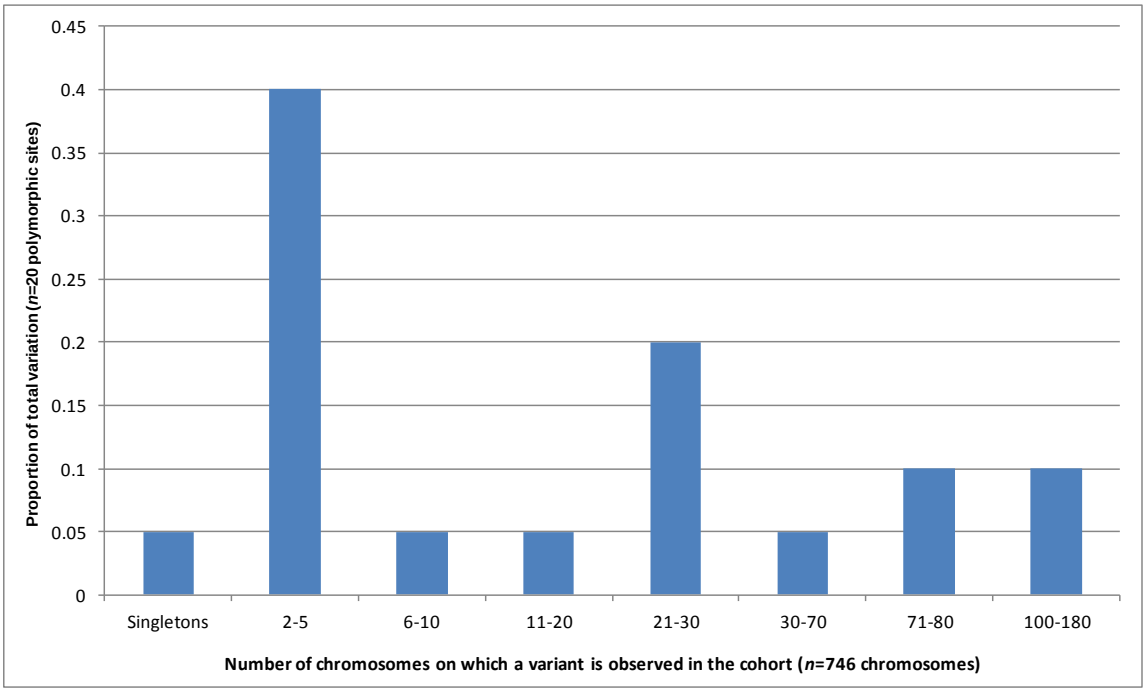
3 exonic polymorphisms were identified in the non-Ethiopian African cohort, compared to 7 in Ethiopians. Of all identified exonic variants, 2 are non-synonymous and only observed in Ethiopians. 40% of all identified polymorphic sites are within intron 3, including the *CYP3A5*3* defining polymorphism. A novel 10 base pair deletion was identified in intron 1 in five heterozygous individuals from West Central Africa. The deletion was not predicted to affect pre-mRNA splicing by BDGP (http://www.fruitfly.org/seq_tools/splice.html). However given the close proximity of this deletion to the intron 1 splice site, experimental evidence will provide an indication of how this deletion may affect *CYP3A5* transcription.

The proportion of singleton variants identified in the eight non-Ethiopian African populations was 0.05% (see Figure 4.3). There were more high and intermediate frequency variants than rare observed from the data.

4.2.2 CYP3A5 diversity in sixteen populations

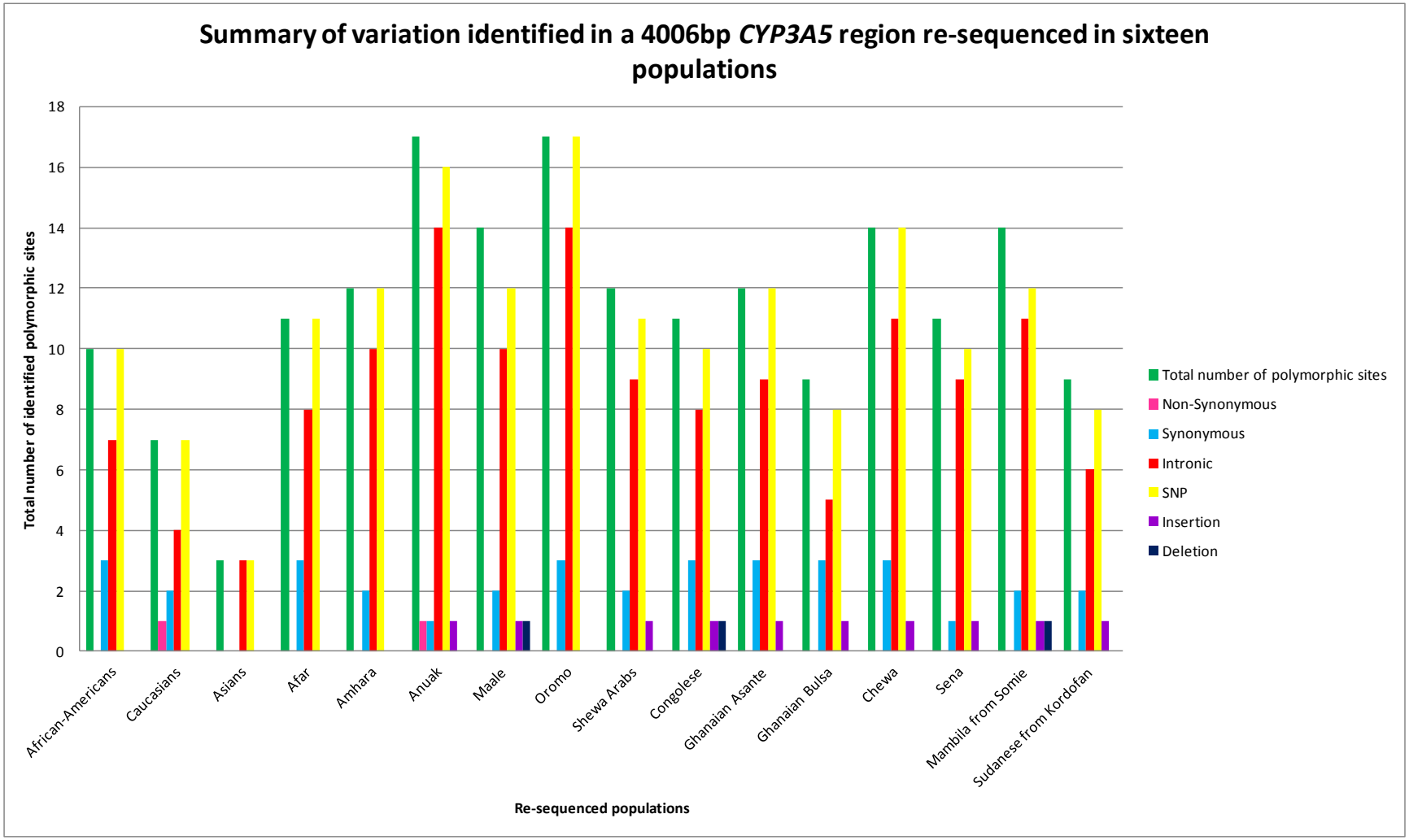
A total of 39 variants were identified within a 4006bp region of *CYP3A5* in all sixteen populations (Figure 4.4). 10 (~25.6%) were exonic polymorphisms; 2 non-synonymous, 1 insertion (*CYP3A5**7), 3 synonymous (one of which defines the *CYP3A5*6* mutation) and 4 in the 5′ UTR of exon 1. All identified non-synonymous polymorphisms, except one in exon 4 identified in a single heterozygous Ethiopian (see Table 5.10), were predicted to have a benign effect on protein function (examined using PolyPhen2). The proportion of exonic polymorphisms observed is higher than neutral expectations for protein coding genes; although the observed number of amino acid changes relative to the number of codons (~0.008%) is lower than reported for protein coding genes (0.56%) (Kitano et al. 2004). No identified exonic polymorphism was predicted to affect mRNA splicing; and no polymorphism occurred in a consensus splice site. The highest frequency variants were *CYP3A5*3, CYP3A5*6* and *CYP3A5*7*.

Figure 4.3: The number of times a particular variant is observed within the non-Ethiopian sub-Saharan African cohort (n=746 chromosomes). The frequency refers to the number of chromosomes on which a particular variant was identified. A "singleton" is a variant that was observed in a single heterozygous individual. The *y*-axis shows the proportion of the total amount of identified variation that is attributed to variants of particular frequencies; i.e. singletons account for ~0.05% of all identified polymorphic sites.



Consistent with previous studies of global diversity, populations with recent African ancestry have more diversity (even for a small 4006bp *CYP3A5* region) than observed in Europeans and Han Chinese (see Figure 4.4). Han Chinese individuals had the lowest number of polymorphic sites. Non-synonymous polymorphisms were only identified in European individuals and in the Anuak. Only populations with recent African ancestry had the *CYP3A5*7* defining T allele insertion in exon 11. Populations with recent African ancestry are also more varied in the types of variation observed; no insertion or deletion was identified in Europeans or Han Chinese which is consistent with comparisons of other genomic regions (Campbell and Tishkoff 2010). Populations with recent African ancestry appear to be homogeneous in the type of variation that is observed. No one population had a higher number of insertions or deletions in the gene region than the others.

Figure 4.4: A graph summarising the identified variation in a 4000bp region of *CYP3A5,* re-sequenced in sixteen populations. The key summarises the types of variation identified and the graph shows the total numbers of each type of polymorphic site identified from the re-sequenced region in each population. The corresponding Table with details of each identified variant and its frequency in the cohort is provided in Supplementary Table 1 (on CD).

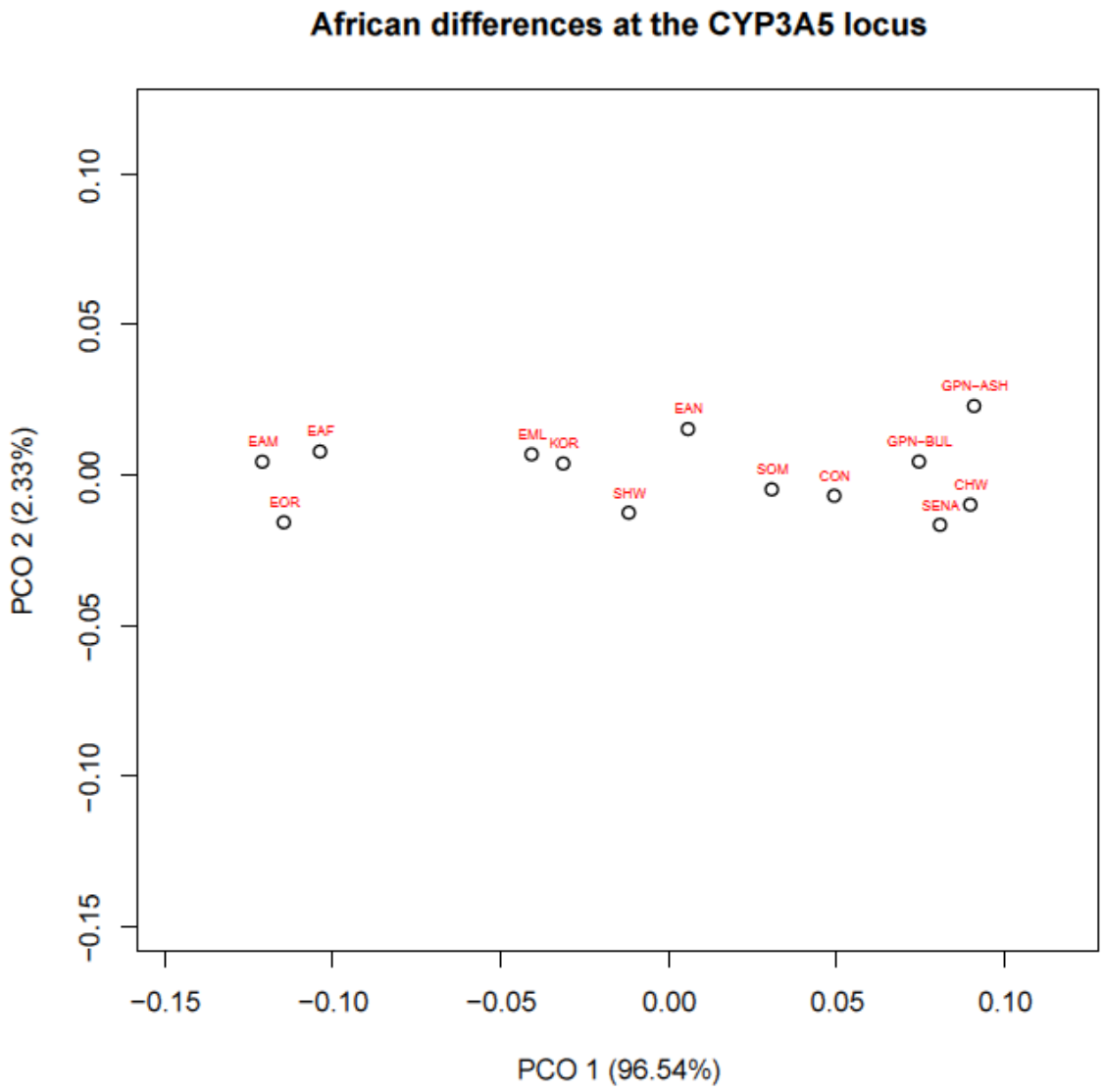


4.2.3 Population structure at the CYP3A5 gene

Allelic data were used to calculate pairwise F_{ST} so that singleton variants could be accounted for when assessing inter-population differences. Pairwise F_{ST} values were used to perform PCO analysis. The results of intra-African analyses (Figure 4.5) are consistent with those reported in chapter 3; the Afar, Amhara and Oromo are distinct from other sub-Saharan African populations. However a comparison of all sixteen groups (Figure 4.6) found that African groups cluster together and Europeans and Asians are the outlying populations.

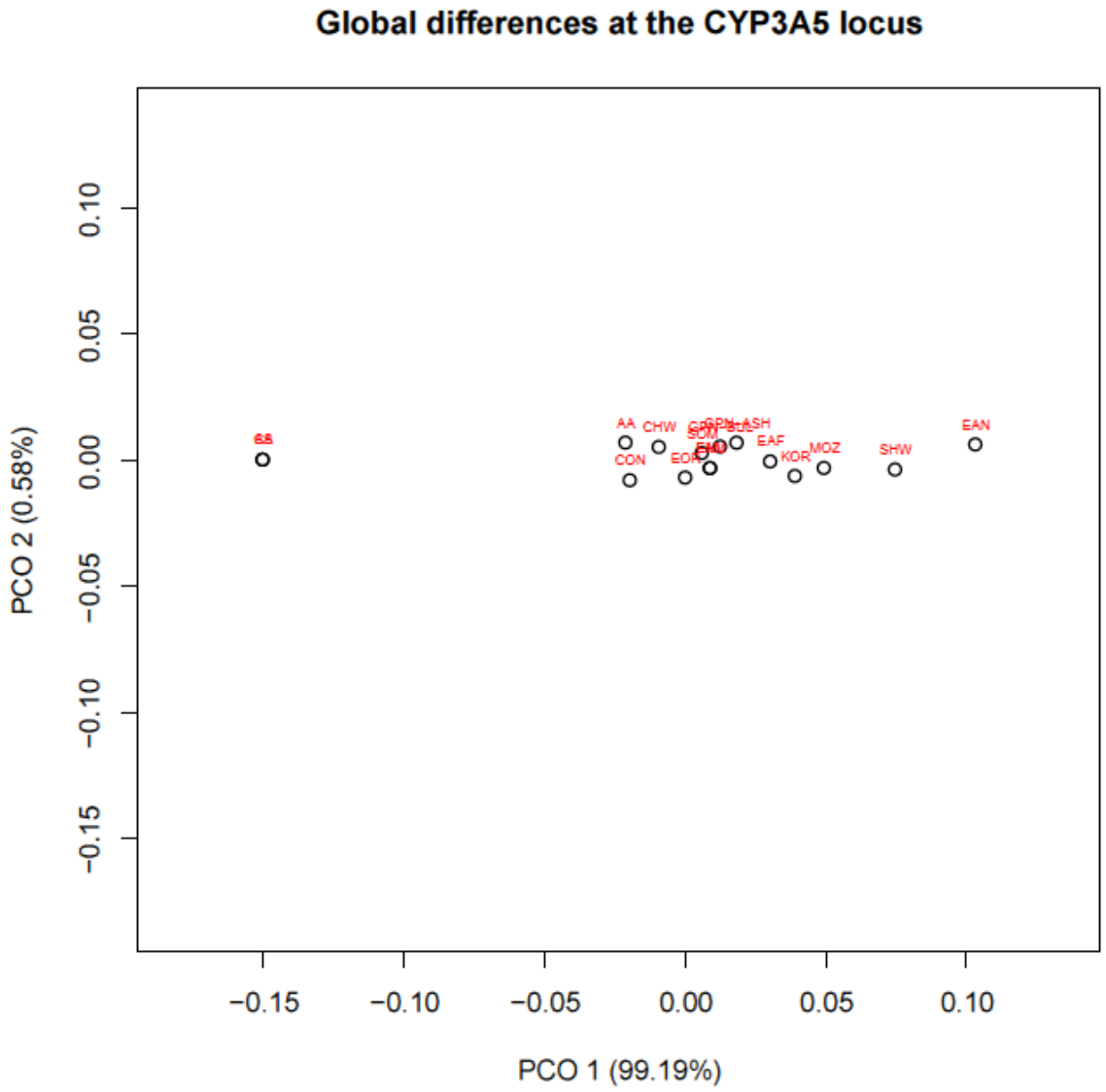
A notable feature of the PCO analyses is that the majority (over 96%) of population differentiation at the *CYP3A5* gene is explained by the first principal component. As reported in chapter 3, inter-population differences in the frequency of the *CYP3A5*3* allele shape population structure at the gene; even when variation over a larger gene region is considered. Few variants were identified in Europeans and Han Chinese other than *CYP3A5*3* which is the main factor which differentiates these populations from Africans at *CYP3A5*. An interesting observation is that *CYP3A5*3* frequencies also shape population structure in Africa, followed by *CYP3A5*6, CYP3A5*7* and all additional variation identified in these populations. It is possible that within a larger genomic region there may be additional variation which will further differentiate African populations. However the data from this chapter suggest that frequencies of these three alleles are the main factors which influence population structure at this gene.

Figure 4.5: A principal co-ordinates (PCO) plot showing the differences between thirteen African groups, based on re-sequencing data for a 4006bp region of *CYP3A5*. Values along either axis represent the total amount of variation captured by each principal-coordinate. The PCO plot was constructed using pairwise F_{ST} comparisons.



Sample set codes: EAF; Afar, EAM; Amhara, EAN; Anuak, EML; Maale, EOR; Oromo, SHW; Shewa Arabs, CON; Congolese, GPN-ASH; Ghanaian Asante, GPN-BUL; Ghanaian Bulsa, CHW; Malawian Chewa, SENA; Mozambicans from Sena, SOM; Cameroonians from Somie, KOR; Sudanese individuals from Kordofan.

Figure 4.6: A principal co-ordinates (PCO) plot showing the differences between sixteen global populations, based on re-sequencing data for a 4006bp region of *CYP3A5*. Values along either axis represent the total amount of variation captured by each principal-coordinate. The PCO plot was constructed using pairwise F_{ST} comparisons.



Sample set codes: EAF; Afar, EAM; Amhara, EAN; Anuak, EML; Maale, EOR; Oromo, SHW; Shewa Arabs, CON; Congolese, GPN-ASH; Ghanaian Asante, GPN-BUL; Ghanaian Bulsa, CHW; Malawian Chewa, SENA; Mozambicans from Sena, SOM; Cameroonians from Somie, KOR; Sudanese individuals from Kordofan, AA; African-Americans, AS; Han Chinese, CA; individuals of recent European ancestry.

4.2.3 Molecular diversity at the CYP3A5 locus

Analyses of molecular diversity were performed using DnaSP software (version 5.0) and are presented in Table 4.3. Nucleotide diversity estimates are low for each of the sixteen groups, compared to the estimate that approximately 1 in every 1000 nucleotides in the human genome are polymorphic (Sachidanandam et al. 2001; Jobling et al. 2004; Rotimi and Jorde 2010) although this may reflect the small size of the *CYP3A5* region analysed. The Afar, Amhara and Oromo have marginally lower estimates of nucleotide diversity compared to other African populations. Tajima's *D* values indicate that there is a skew towards rare variants in East Africa and in four other sub-Saharan African groups; although significant departures from neutrality were not observed in the Afar following Bonferonni correction for multiple tests ($0.004<p<0.05$). A skew towards rare variants in East Africa may indicate evidence of differential selective pressures across the African continent. Although it is important to note that the results reported here are for a smaller region than that analysed in chapter 6. Comparative analyses of the entire *CYP3A5* gene region between the thirteen groups will identify whether there is evidence of differential selective pressures across the African continent.

A skew towards rare variants was observed in all three Coriell populations. Significant departures from neutrality were not observed in Han Chinese individuals; however this may be due to the paucity of *CYP3A5* variation observed within this population. The Afar and Oromo have a greater skew towards rare variants than Han Chinese and African-Americans; although this reflects the paucity of variation observed in the latter two groups. A significant departure from neutrality was observed in both Europeans and African-Americans, however the departure was not significant for Europeans following Bonferonni correction ($0.003<p<0.004$).

Table 4.3: Molecular diversity estimates for sixteen global sample sets. Statistically significant departures from neutrality, following Bonferonni correction are highlighted in green (correction for 16 tests; adjusted significance level is $p \leq 0.003$).

	AA	AS	CA	EAF	EAM	EAN	EML	EOR	SHW	CON	GPN-ASH	GPN-BUL	CHW	SENA	SOM	KOR
Number of polymorphic sites	9	2	5	12	12	17	14	17	12	11	12	8	14	11	14	9
Number of singletons	2	1	1	1	0	1	2	1	0	0	0	0	0	0	1	0
Nucleotide diversity (π)	4.3×10^{-4}	1.0×10^{-4}	1.6×10^{-4}	3.0×10^{-4}	3.2×10^{-4}	4.3×10^{-4}	4.1×10^{-4}	2.2×10^{-4}	5.2×10^{-4}	4.3×10^{-4}	3.9×10^{-4}	4.3×10^{-4}	7.1×10^{-4}	4.7×10^{-4}	4.9×10^{-4}	4.8×10^{-4}
Tajima's *D*	-0.918	-0.932	-1.603	-1.15	-0.967	-1.33	-0.764	-1.91	1.44	-0.701	-0.584	-0.0398	-0.680	1.04	0.552	0.168
Fu and Li's *D**	-0.812	-1.680	-1.846	-0.301	0.106	-1.59	0.778	-0.177	0.805	1.25	0.05	1.25	1.12	0.825	0.919	-0.144
Fu and Li's *F**	-0.997	-1.69	-2.075	-0.716	-0.342	-1.79	0.255	-0.992	1.18	0.69	-0.177	0.99	0.616	1.05	0.943	-0.055
Fu's *FS*	**-12.03**	-0.218	5.18	-2.37	**-7.99**	**-11.6**	**-7.85**	**-14.3**	-0.108	-2.54	-1.50	-1.44	-1.94	0.235	0.535	-1.089

Sample set codes: AA; African-Americans, AS; Han Chinese, CA; individuals of recent European ancestry, EAF; Afar, EAM; Amhara, EAN; Anuak, EML; Maale, EOR; Oromo, SHW; Shewa Arabs, CON; Congolese, GPN-ASH; Ghanaian Asante, GPN-BUL; Ghanaian Bulsa, CHW; Malawian Chewa, SENA; Mozambicans from Sena, SOM; Cameroonians from Somie, KOR; Sudanese individuals from Kordofan.

4.3 Statistical association of identified variation

4.3.1 Haplotype inference

A total of 35 haplotypes were inferred for a 4006bp region from data for sixteen populations; 29 were observed in Africa. ~42.9% of all haplotypes were *CYP3A5*1,* ~34.3% were defined by the *CYP3A5*3* mutation alone, ~14.3% by *CYP3A5*6*, ~5.7% by *CYP3A5*7* and a single recombinant *CYP3A5*3/CYP3A5*6* haplotype was observed in 9 heterozygous individuals; see Figures 4.7a-b. *CYP3A5*1* was the most diverse haplogroup; followed by *CYP3A5*3*; *CYP3A5*6*; *CYP3A5*7* and *CYP3A5*3/*6* (Figure 4.7c).

Within Africa, *CYP3A5*3* haplotype frequencies for Shewa Arabs and Sudanese individuals from Kordofan were comparable to the Ethiopian Maale. ~40% of inferred haplotypes for Shewa Arabs and ~43% of those for Sudanese individuals from Kordofan were defined by the *CYP3A5*3* variant. *CYP3A5*7* was observed at high frequencies in non Ethiopian sub-Saharan Africans; the highest frequencies of the haplotype were observed in the Chewa from Malawi (frequency ~22%).

Within Africa, variation in haplotype frequencies and distribution is highest in populations outside of Ethiopia, see Figure 4.7b. Consistent with Y chromosome and mitochondrial DNA (de Filippo et al.), Niger-Congo speakers are homogeneous in their *CYP3A5* haplotype frequencies (Figures 4.7b). This is consistent with the observed correlation between major language family and *CYP3A5* structure reported in chapter 3. Population similarities between Niger-Congo speaking groups, who are spread over large geographic distances, are due to recent demographic events. Haplotype patterns in the Nilo-Saharan speaking Sudanese and Afro-Asiatic speaking Shewa Arabs differ from Niger-Congo groups. Shewa Arabs and Kordofanian Sudanese also have the highest frequencies of *CYP3A5*3* haplotypes and appear to have more variation within this haplotype class than Niger-Congo speakers.

The majority of haplotypes observed in Han Chinese and European populations were defined by *CYP3A5*3*. Of the three Coriell groups, haplotype diversity was greatest in African-Americans; although fewer haplotypes were observed in this population than in the collective African cohort.

Figure 4.7a: The 35 global haplotypes inferred from re-sequencing data for a 4006 base pair region of CYP3A5. Each position is an identified polymorphic site. Positions are numbered from the ATG start codon where base A is +1; and correspond to details provided in chapter 4. Yellow indicates the ancestral allele at each position (as inferred from chimpanzee sequence) and blue the derived (polymorphism). "N" refers to the total number of chromosomes of a particular haplotype; specific frequencies per population are provided in Figure 4.5b.

Haplotype class	Code	-86	-74	-15	74	127	136	182	289	318	5209	5229	5244	5416	6980	7201	7354	7355	14684	14714	14830	14877	26943	27044	27128	N
*1	1	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	400
	2	G	C	A	I	**A**	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	2
	3	G	C	A	I	**A**	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	**G**	G	A	-	1
	4	G	C	A	I	G	C	**A**	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	2
	5	G	C	A	I	G	C	C	**C**	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	3
	6	G	C	A	I	G	C	C	G	G	**T**	G	C	C	A	C	T	C	G	A	C	A	G	A	-	41
	7	G	C	A	I	G	C	C	G	G	C	**A**	C	C	A	C	T	C	G	A	C	A	G	A	-	7
	8	G	C	A	I	G	C	C	G	G	C	G	C	C	A	**T**	T	C	G	A	C	A	G	A	-	86
	9	G	C	A	I	G	C	C	G	G	C	G	C	C	A	**T**	T	C	G	A	C	**G**	G	A	-	6
	10	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	**C**	C	G	A	C	A	G	A	-	5
	11	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	**G**	C	A	G	A	-	21
	12	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	**T**	A	G	A	-	3
	13	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	**G**	G	A	-	23
	14	G	C	A	I	G	C	C	G	**A**	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	7
	15	G	C	**C**	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	1
*3	16	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	608
	17	G	**T**	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	20
	18	G	C	A	**D**	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	5
	19	G	C	A	I	G	C	C	G	G	**T**	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	33
	20	G	C	A	I	G	C	C	G	G	**T**	G	C	C	**G**	C	T	**T**	G	A	C	A	G	A	-	1
	21	G	C	A	I	G	C	C	G	G	C	G	**T**	C	**G**	C	T	C	G	A	C	A	G	A	-	2
	22	G	C	A	I	G	C	C	G	G	C	G	C	**T**	**G**	C	T	C	G	A	C	A	G	A	-	2
	23	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	**C**	C	G	A	C	A	G	A	-	1
	24	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	**T**	G	A	C	A	G	A	-	2
	25	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	**G**	-	5
	26	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	**A**	A	-	1
	27	**A**	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	1
*6	28	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	256
	29	G	C	A	I	G	C	C	G	G	**T**	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	4
	30	G	C	A	I	G	C	C	G	G	C	G	C	C	A	**T**	T	C	**A**	A	C	A	G	A	-	3
	31	G	C	A	I	G	**T**	C	G	G	C	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	1
	32	G	C	A	I	G	C	C	G	**A**	C	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	2
*3/*6	33	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	**A**	A	C	A	G	A	-	9
*7	34	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	**T**	76
	35	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	**C**	C	G	A	C	A	G	A	**T**	1

Figure 4.7b: The frequencies of each inferred haplotype (as shown in Figure 4.5a) in each global population. Haplotype codes correspond to those listed in Figure 4.7a. The haplotype class, to which each haplotype belongs, is shown in brackets following the entry in the key.

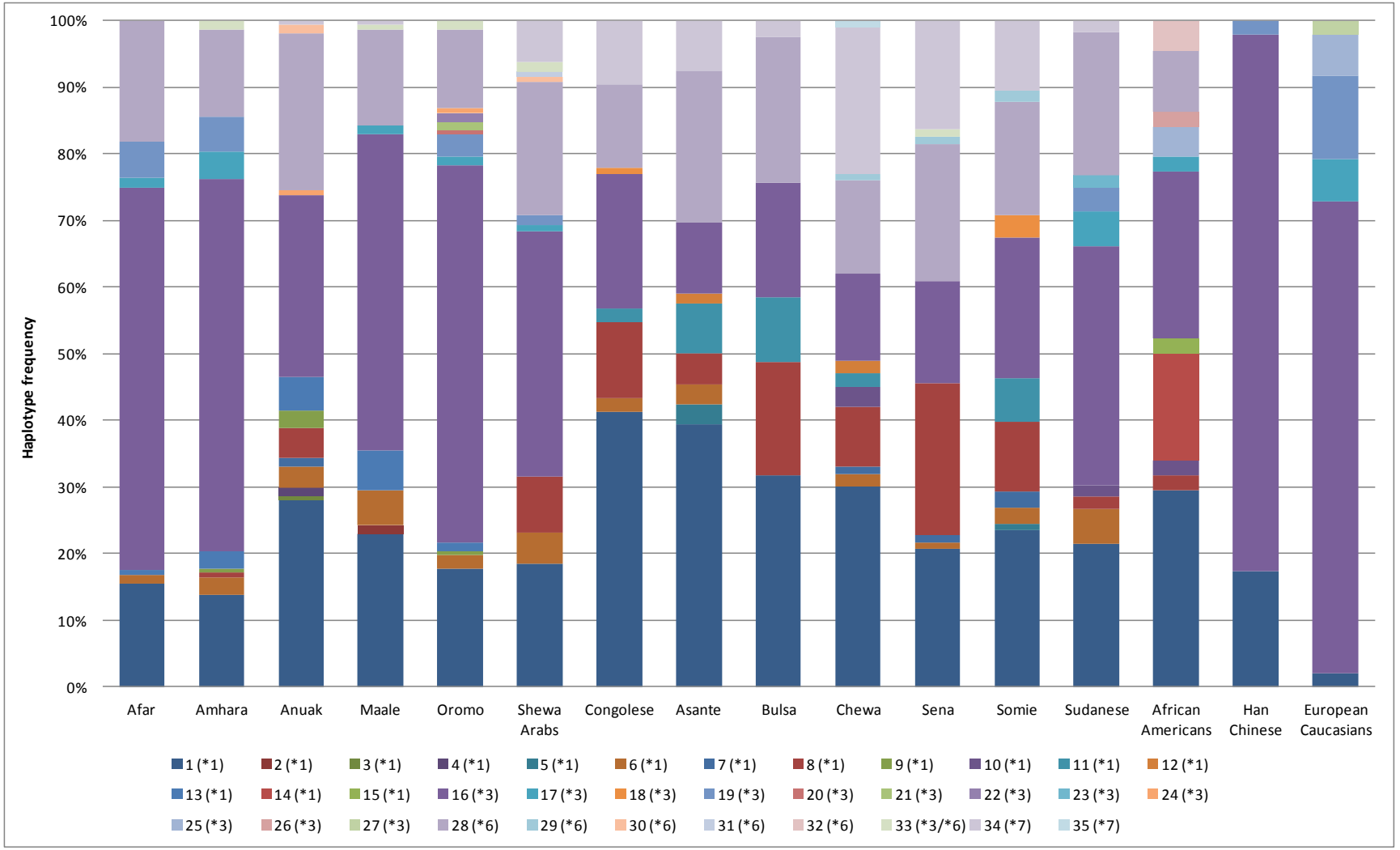
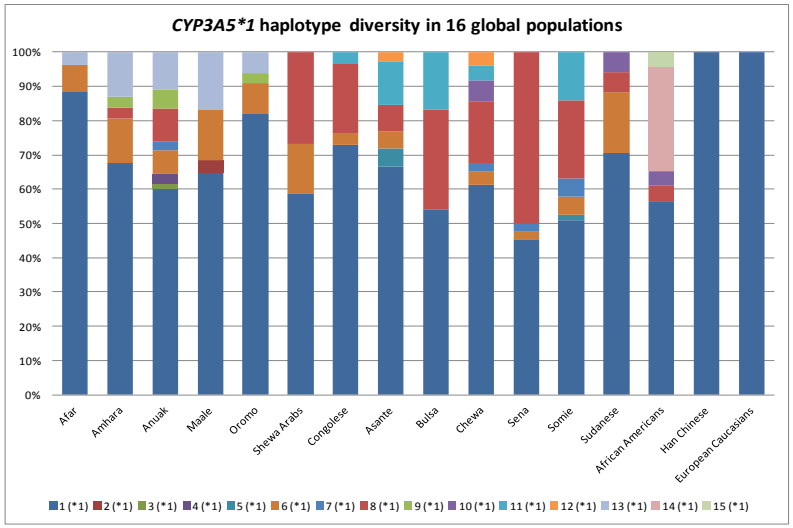
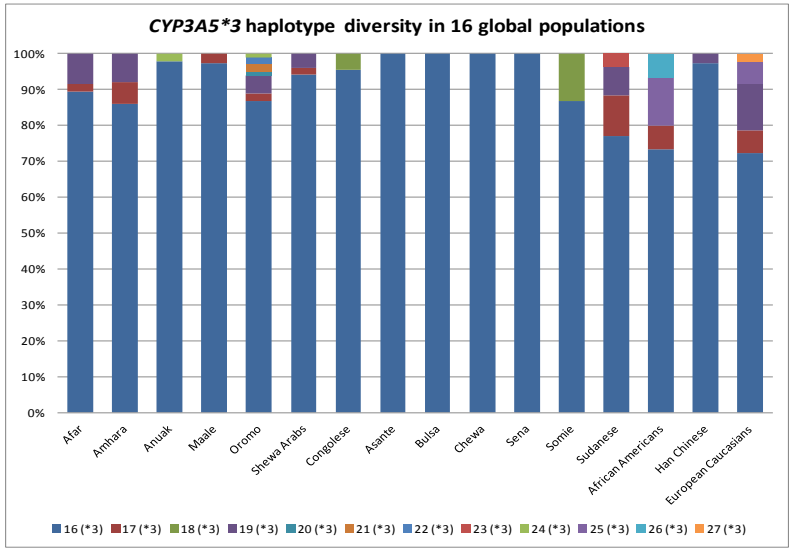
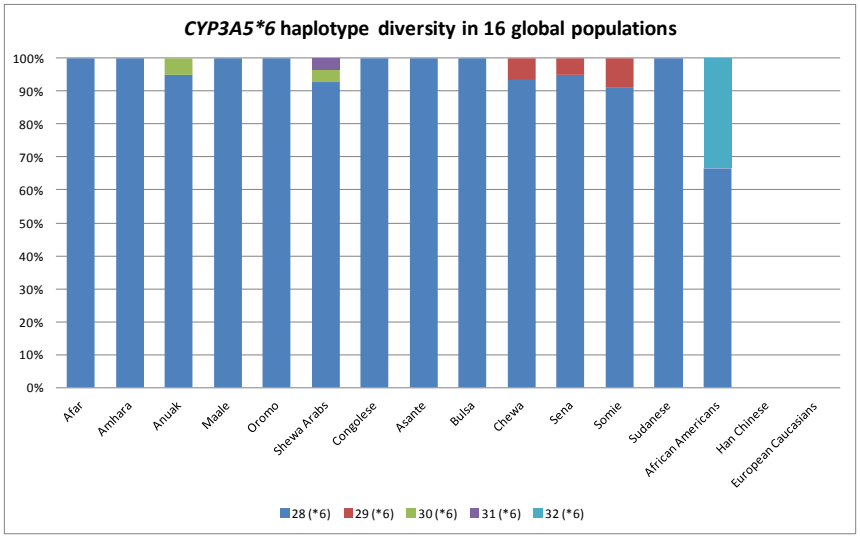
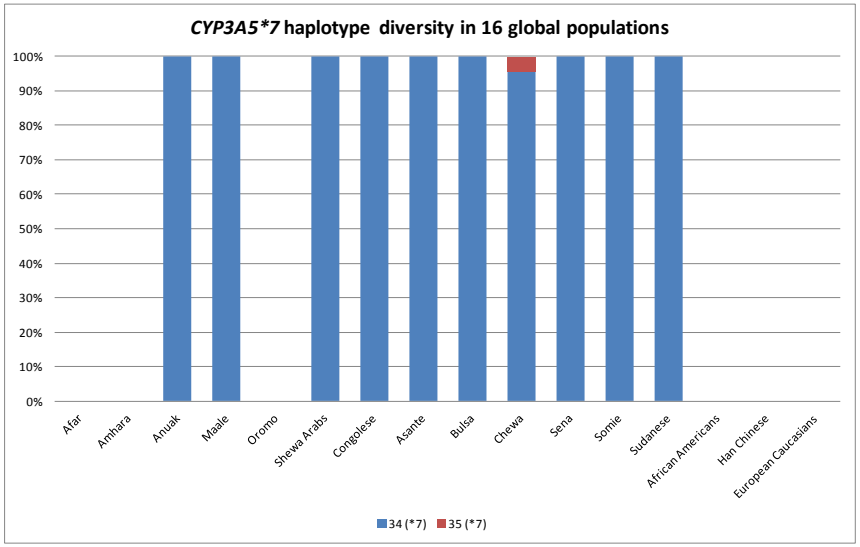


Figure 4.7c: Global diversity in each *CYP3A5* haplogroup; the key corresponds to haplotypes (numbered as in Figure 4.5a).









4.3.2 Assessing the diversity of inferred haplotypes

Nei's gene diversity (*h*) was calculated to assess haplotype diversity; between populations and *CYP3A5* haplotype classes. The significance of inter-population differences in gene diversity was assessed using an exact test of population differentiation (presented in Table 4.4).

Within Africa gene diversity was highest in groups outside of Ethiopia and North Africa (Kordofanians from Sudan), see Figure 4.8. This is expected given their comparatively higher frequencies of *CYP3A5*3* haplotypes than observed in other African populations. *CYP3A5*3* haplotypes have less diversity than those defined by the ancestral *CYP3A5*1* allele (Figure 4.7). Although from Figure 4.9, it is also apparent that *CYP3A5*3* haplotype diversity is highest in populations with high frequencies of the allele, and in African-Americans and Sudanese Kordofanians *CYP3A5*3* haplotype diversity is almost equal to that seen in the *CYP3A5*1* haplogroup. Within the sub-Saharan African cohort, the Afar, Amhara, Oromo and Somie have the highest frequencies of the *CYP3A5*3* allele and subsequent higher levels of diversity in the *CYP3A5*3* haplogroup.

Figure 4.8: Nei's h estimate of gene diversity, for a 4006bp *CYP3A5* region, in sixteen populations. Error bars denote standard deviation, calculated as outlined in chapter 2, and bars are coloured according to the geographic region each population belongs to.

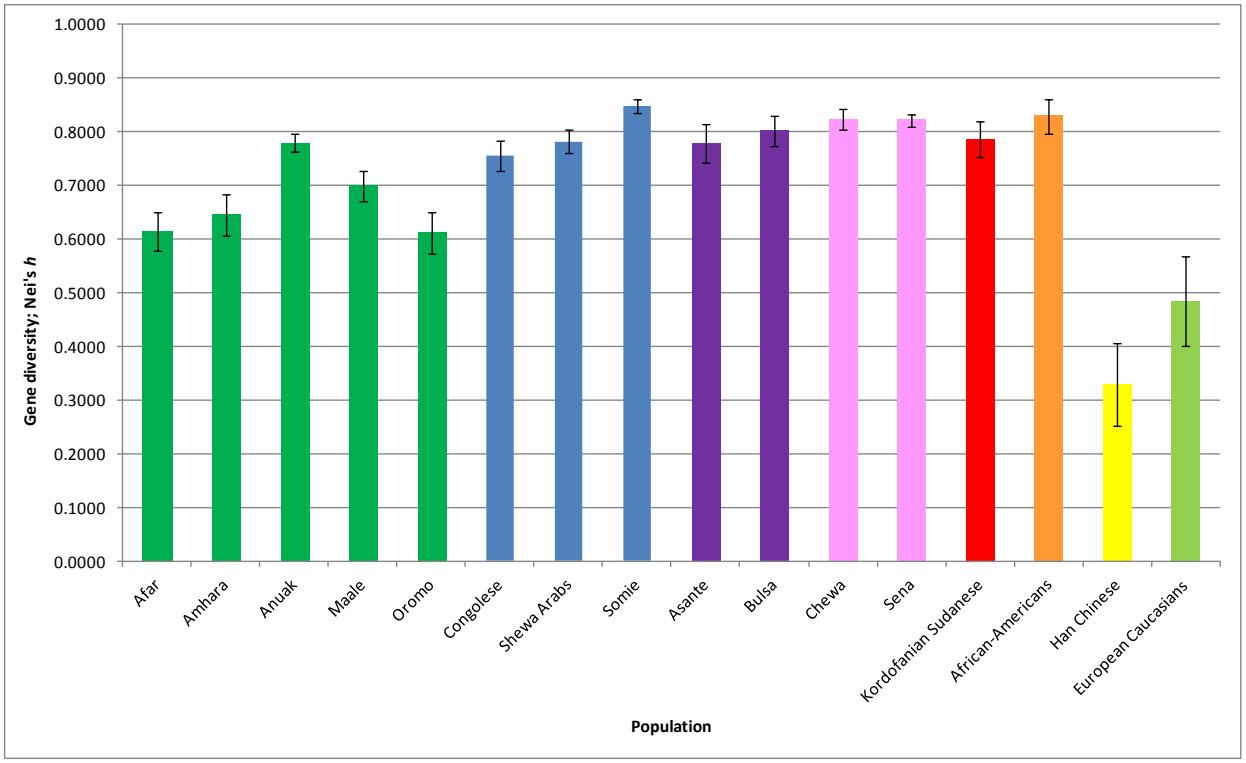
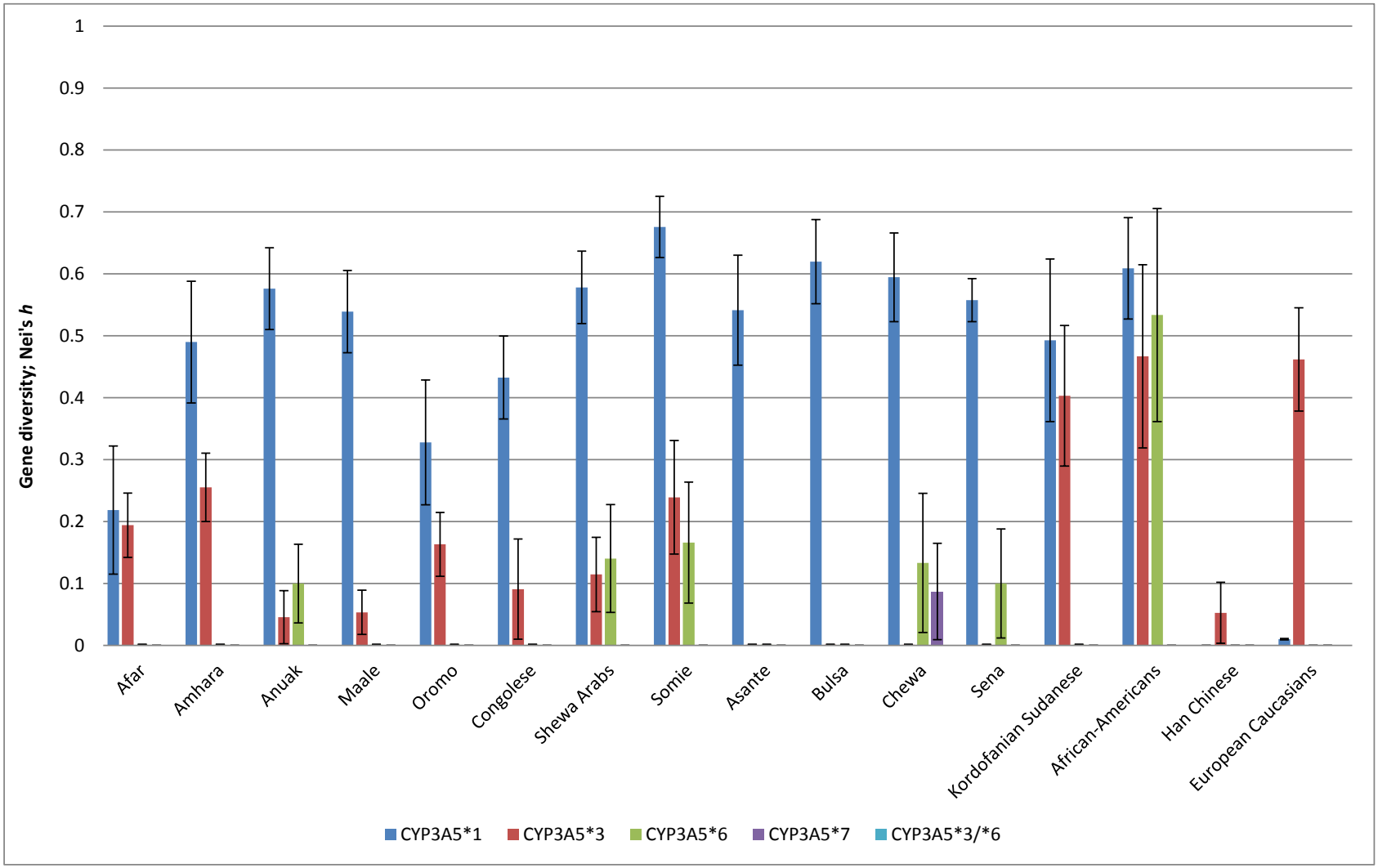


Table 4.4: An exact test of population differentiation to measure the significance of differences in gene diversity estimates (measured by Nei's *h*) in sixteen populations. Statistically significant differences following Bonferonni correction (for 16 tests; adjusted *p*-value = 0.003125) are shown in bold and highlighted in green. Population codes are as provided for Figures 4.3-4.4.

	EAF	EAM	EAN	EML	EOR	SHW	CON	ASH	BUL	CHW	SENA	SOM	KOR	AA	AS	CA
EAF	*															
EAM	0.42537	*														
EAN	<0.0001	<0.0001	*													
EML	0.00004	0.00336	0.0001	*												
EOR	0.4755	0.65113	<0.0001	0.01571	*											
SHW	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	*										
CON	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.00055	*									
ASH	<0.0001	<0.0001	0.00007	<0.0001	<0.0001	<0.0001	0.05376	*								
BUL	<0.0001	<0.0001	0.00618	<0.0001	<0.0001	0.00553	0.15141	0.32365	*							
CHW	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.06596	0.04871	0.03493	*						
SENA	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.00006	0.00574	0.00001	0.02583	0.05286	*					
SOM	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0001	0.11521	0.08024	0.60052	0.03531	0.01705	*				
KOR	0.00558	0.04739	0.02513	0.00942	0.04052	0.21083	0.00022	0.0007	0.00252	<0.0001	<0.0001	0.00021	*			
AA	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.00013	0.00097	<0.0001	<0.0001	<0.0001	0.0057	*		
AS	0.00455	0.03955	<0.0001	0.00071	0.29675	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	*	
CA	<0.0001	0.00024	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.00229	*

Figure 4.9: Nei's h estimates for each of five *CYP3A5* haplotype classes (*CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*7* and *CYP3A5*3/*6*) in each re-sequenced population. Error bars denote standard deviation.



4.4 Testing for correlations between *CYP3A5* diversity, environmental and demographic factors

Examining population differences using F_{ST} and pairwise exact tests of population differentiation alone cannot differentiate between variables which influence genetic diversity and similarity; for example populations which are geographically close together tend to be much more similar than those located further away. However, sometimes populations located closely to each other are found to differ; perhaps due to specific environmental boundaries which may prevent gene flow (such as the Sahara desert between North and sub-Saharan Africa) (Jobling et al. 2004). In these cases the true geographic distance between populations is larger than first thought. These differences can be modelled using spatial autocorrelation analyses; however the samples used in this thesis were largely collected from multiple ethnic groups in one location within the specified country meaning that detailed spatial autocorrelation analyses cannot be performed. However in the absence of data on specific environmental factors, a Mantel test is often useful (Mantel 1967). Often information on pairwise comparisons of different aspects of the data is available as matrices; a Mantel test examines whether there are correlations between two or more matrices for a given population. Such comparisons may highlight specific differences which help to explain the observed genetic differences. There is a problem with Mantel tests in that each pairwise correlation between matrices is not strictly independent. However, Mantel tests have been critical in improving understanding of the origins of different populations, such as in Europe (Jobling et al. 2004).

Although the re-sequencing data examined in this chapter are for a small region of *CYP3A5*, comparisons of matrices may highlight specific environmental variables which are shaping the differences between the populations. The objectives of the analyses presented in this section are to examine whether there are correlations between geographic proximity and genetic similarity; and to examine whether there is a correlation between *CYP3A5*3* haplotype diversity and latitude. Of the sixteen populations geographic co-ordinates were available for all thirteen African populations; the Coriell datasets were excluded from these analyses.

A Mantel test to compare pairwise genetic differences (measured by F_{ST}) and geographic proximity (in kilometres) was performed in the R-programming environment. The results were significant (p=0.0064); meaning that there is a correlation between geographic proximity and genetic similarity at the *CYP3A5* locus in Africa.

As a strong positive correlation between *CYP3A5*1/*3* allele frequencies and latitude was identified in chapter 3, a test for the correlation (if any) between *CYP3A5*1* and *CYP3A5*3* haplotype diversity and latitude was performed by Spearmann's Rank Correlation analyses. No significant correlation between gene diversity in the *CYP3A5*1* haplogroup (Rho = 0.5536, *p*=0.1813) or *CYP3A5*3* haplogroup (Rho = -0.1142, *p*=0.7102) and latitude was observed. It is much more likely that haplogroup diversity is correlated with increased frequencies of specific *CYP3A5* alleles.

4.5 Discussion

The results presented in this chapter are an extension of those presented in chapter 3. Population structure at the *CYP3A5* gene is almost entirely influenced by frequencies of the *CYP3A5*3* mutation. This is perhaps surprising given that Africans have high levels of genetic diversity; although additional variants do affect intra-African population structure at the gene. *CYP3A5* haplotypes defined by the low/non-expresser *CYP3A5* alleles are comparatively less diverse than for those defined by *CYP3A5*1.* This is expected as *CYP3A5*1* is the ancestral allele. A significant reduction in diversity levels on low/non-expresser haplogroups, from neutral expectations, may be consistent with the alleles being recent mutations; rapid population expansion; or positive selection for low/non-expression. Significant departures from neutrality can be assessed by comparing larger genomic regions, which include the *CYP3A5* gene, in multiple populations.

4.5.1 Intra-African diversity at the CYP3A5 locus

Consistent with the results from the geographic survey (chapter 3) there is considerable heterogeneity within East Africa. Overall diversity in the gene region was low for all populations and the intra-African structuring seen from PCO analysis is similar to that in Figure 3.3. The Afar, Amhara and Oromo (all Afro-Asiatic speaking) populations differ from other sub-Saharan African groups. Kordofanian Sudanese (North Africans) also differ significantly from other sub-Saharan Africans; consistent with the Sahara desert acting as a major barrier to gene flow across the continent (Cruciani et al. 2002). Interestingly, no East African population differed from Kordofanian Sudanese. It is possible that the short geographic distance between these populations facilitates gene flow readily. It is also possible that North African admixture with East Africans, coupled with gene flow from the Arabian Peninsula, has influenced the patterns of *CYP3A5* diversity that are observed in the region (Richards et al. 2003).

The finding that population differentiation, as a result of *CYP3A5* variation, is largely explained by differences in *CYP3A5*3* allele frequencies is consistent with a hypothesis of selection on the allele in populations outside of Africa. There are two feasible explanations for high frequencies of *CYP3A5*3,* coupled with a paucity of variation on *CYP3A5*3* haplotypes and a skew towards rare variants in three of the five Ethiopian populations. The first is that Ethiopian heterogeneity is a result of differential selective pressures on the gene across the continent. This would mean that there may be specific environmental pressures which influence the high levels of diversity; even between East African populations. An alternative explanation is that East African heterogeneity, and the clustering of the Afar, Amhara and Oromo (from PCO analysis) between a larger African cohort and non-African populations is due to admixture with North Africans and populations from the Arabian Peninsula. There is a known genetic contribution of Arabian populations to populations in the North and North East of Ethiopia (Cruciani et al. 2002; Lovell et al. 2005). A comparison with sequences from the Arabian Peninsula will aid in elucidating the proportion of genetic differentiation in East Africa that is attributable to admixture.

A limitation of the *CYP3A5* data presented in this chapter is that the full gene has not been re-sequenced in all 13 African populations. Identifying and analysing variation across the entire gene in these groups may identify more variants which are likely to affect protein function and/or expression. Analyses of the frequencies of identified African *CYP3A5* variants will provide evidence of whether intra-African diversity is a result of differences in demographic history or selective constraints. A comprehensive understanding of African *CYP3A5* diversity will be essential for determining suitable drug concentrations within the region and in predicting disease risks; such as hypertension and hyponatremia.

4.5.2 Inter-population CYP3A5 diversity

A comparison of African *CYP3A5* re-sequencing data with Coriell populations found that *CYP3A5* diversity was higher in populations with recent African ancestry than European and Han Chinese individuals. Many previous studies on *CYP3A5* variation have extrapolated data obtained for African-American populations to predict the patterns of variation and frequencies of CYP3A5 expressers within Africa. However given the extensive amount of variation observed across the African continent, and reported for *CYP3A5* in this thesis, this method of extrapolating data for African-Americans is unlikely to account for East African heterogeneity, such as differences between North and sub-Saharan Africans; or between East

Africans and other sub-Saharan African groups. In order to characterise intra-African *CYP3A5* diversity multiple populations from within the continent must be considered.

Using data from the sixteen populations core haplogroups were identified; those defined by one of the *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles, those defined by the ancestral *CYP3A5*1* and a single recombinant haplotype *CYP3A5*3/*6*. It is possible that there are additional variants which define core *CYP3A5* African haplotypes, outside of the re-sequenced region. Within each haplogroup the modal haplotypes are defined by a paucity of variation. *CYP3A5*3* haplogroup diversity was greatest in Europeans and Han Chinese; populations in which it is believed that a selective sweep of *CYP3A5*3* has occurred (Thompson et al. 2004). Although the frequencies of additional low/non-expresser haplotypes is rare; suggesting that differentiation of the *CYP3A5*3* haplogroup is rare even in populations where the *CYP3A5*3* allele is observed at high frequencies.

One method of establishing whether there is strong evidence of differential selection in Africa is to model expectations for multiple polymorphic sites and compare the expected diversity, under a hypothesis of neutrality, with the observed data. Such a method is similar to simulating datasets under specific conditions (see chapter 7) but simpler as it is less computer-intensive. Additionally such a method does not generate an additional dataset but provide an approximate distribution of allele frequencies within a population. One way of doing this is through the Bayenv program (Coop et al. 2010) which uses Bayesian methods to simulate an expected distribution of allele frequencies, given the data, and then compares the data with those observed to determine the significance of differences between observed and expected data differences. This could be used for comparative analyses of a larger region of the *CYP3A5* gene in African and other global populations to identify populations which may have undergone a selective sweep and/or differential selective pressures within a global cohort.

4.5.3 Intra-African diversity at the CYP3A5 locus is likely to have implications for healthcare of populations within and from the region

One of the most significant findings of this, and the previous, chapter is that Ethiopians differ from other sub-Saharan African populations in the amount of clinically relevant *CYP3A5* variation. An appreciation of how demography, evolutionary history and anthropology shape genetic diversity in medically important genes can help to identify subsets of individuals from a large geographic region who are distinct from a wider patient population. The separation of some East African populations from a wider African cohort will almost certainly have

important medical implications across the continent and challenge the practise of treating "Africans" as a large homogeneous group in medicine.

Within the 4448bp region, in Africans there is a paucity of intermediate or high frequency variants likely to affect CYP3A5 function, other than *CYP3A5*3, CYP3A5*6* and *CYP3A5*7*. This finding complements those from the previous chapter; genotyping of common, clinically important variants, as opposed to re-sequencing of large regions of medically important genes, may be a cost-effective way to identify patients who require idiosyncratic health interventions. However, a novel rare ten base pair deletion immediately adjacent to the 3′ splice site, was identified in five heterozygous individuals and may affect mRNA splicing and consequently protein expression. This deletion occurs on a *CYP3A5*1* haplotype background. Therefore conclusions about an individual's protein expression status based on *CYP3A5*1* genotype may overlook rare, clinically relevant variation. The ten base pair deletion may be an as yet uncharacterised clinically important African *CYP3A5* mutation and suggests that an appreciation of the effects of rare mutations in the gene may be necessary in individuals with recent African ancestry to help prevent adverse clinical outcomes. Experimental techniques will help to establish the effect of such a large deletion in the gene sequence on protein expression (discussed in section 5.4.2). The 4448bp region analysed in this chapter is also approximately half of the gene, and it is likely that additional high-, intermediate- and low-frequency *CYP3A5* variants which are likely to affect CYP3A5 expression and inter-population differences in clinical outcomes were overlooked. The following chapter will examine *CYP3A5* variation in more detail. Bioinformatics analyses of *CYP3A5* variation identified in a 4448bp region (reported in this chapter) and in the entire gene (re-sequenced in Ethiopian populations) will identify which, if any, variants are predicted to impact CYP3A5 expression by impacting the gene at the level of transcription or translation.

The World Health Organisation (WHO) has data on hypertension incidence and prevalence for multiple populations; including Africans. However, the data have not been standardised by age and geographic co-ordinate information is not available for each individual population (personal correspondence with Dr Richard Cooper[3] and Dr Charles Rotimi[4]). This makes the data difficult to interpret and compare accurately. There are groups, such as those run by Dr Cooper and Dr Charles Rotimi, who look at incidence of hypertension

[3] Dr Richard Cooper is a physician epidemiologist with an interest in cardiovascular disease based in the Department of Preventative Medicine and Epidemiology at Loyola University Medical Centre in Chicago.

[4] Dr Charles Rotimi is the director for the Centre for Research on Genomics and Global Health at the National Institutes of Health in Maryland, USA. His research is focused on examining hypertension risk factors in the African Diaspora

within the African Diaspora however the data are still being collated and it was not possible to incorporate them into the analyses presented in this chapter. Data on differential prevalence of hypertension in individuals with recent African ancestry who reside at high latitudes, including diverse Ethiopian groups, may identify intra-African differences in the susceptibility to and incidence of hypertension.

Conversely if *CYP3A5*3* confers a disadvantage in populations closest to the equator (as previously reported (Thompson et al. 2004)) then high frequencies of *CYP3A5*3* and other, novel low/non-expresser variants observed in the Afar, Amhara and Oromo from Ethiopia (see chapter 6) would be expected to have a detrimental effect; given their geographic proximity to the equator, and mean that they have a higher predisposition to conditions such as hyponatremia (not enough sodium retention). Hyponatremia has, to date, not been considered as a consequence of *CYP3A5* variability because Ethiopian populations have been largely underrepresented in large genotyping and *CYP3A5* re-sequencing surveys. However, the data from this chapter, and subsequent chapters, suggests that there may be a gradient within the country regarding disease susceptibility.

Aside from hypertension and hyponatremia risk, there are additional implications for healthcare within the region as a result of *CYP3A5* variability. CYP3A5 metabolises many therapeutic drugs used to treat a wide spectrum of diseases endemic in Africa (Fellay et al. 2005; Diczfalusy et al. 2008). The African cohort analysed in this chapter are at risk of multiple communicable and non-communicable diseases (Mabayoje 1956; Aspray et al. 1998; Coleman 1998; Walker et al. 1998; Wurthwein et al. 2001). Given the paucity of funding for many diseases, particularly communicable, within the region there has been a shift in focus towards mass drug administration regimens to manage infections within the continent (Hotez 2009; Smits 2009). However the current emphasis on mass drug administration overlooks population-specific differences in drug treatment responses and may bias against, and so be detrimental to, specific populations within/from Africa. What is unknown at present is whether there are incidences of adverse clinical outcomes associated with mass drug administration campaigns which occur in patients as a result of CYP3A5 variability (or that of any other drug metabolising enzymes). Critically, the distribution of drugs as part of large campaigns at European specific doses overlooks intra-African diversity in medically important genes and may increase the incidence and risk of CYP3A5-associated adverse clinical outcomes.

5. Assessing the functional significance of *CYP3A5* variation in Africa

5.1 Chapter overview and aims

Significant intra-African diversity in the frequencies of functionally important *CYP3A5* alleles is likely to influence population differences in the susceptibility to adverse clinical outcomes attributed to variability in CYP3A5 expression (discussed in section 1.2.4). In chapters 3 and 4 it was reported that the low/non-expresser *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles occur largely on independent haplotype backgrounds; suggesting that traits causing low/non-expression of CYP3A5 have evolved more than once. It is possible that there are additional variants, which occur on the same or different haplotype backgrounds as *CYP3A5*3, CYP3A5*6* or *CYP3A5*7,* and contribute to polymorphic enzyme expression in global populations. Additional *CYP3A5* variation may also be responsible for adverse clinical outcomes in patients with recent African ancestry despite having a *CYP3A5*1* genotype.

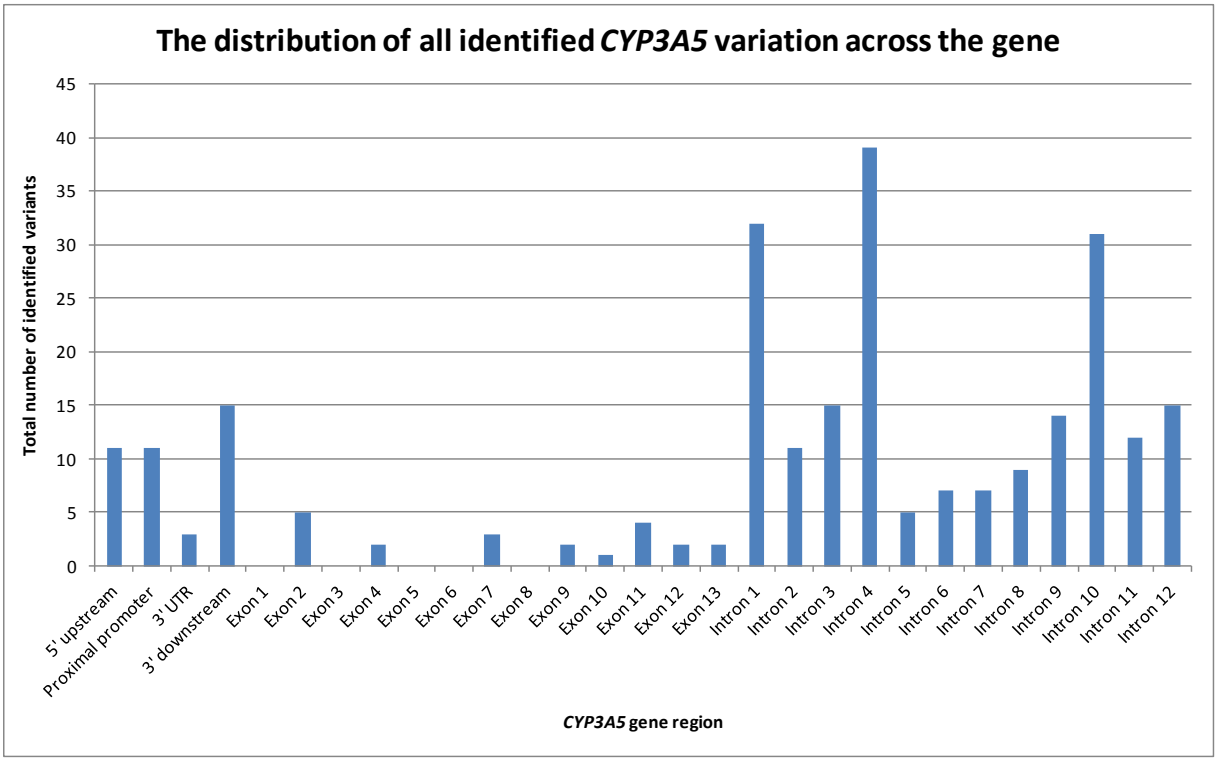
This chapter aims to use bioinformatics tools to analyse all variation identified in a 12,237 base pair region of *CYP3A5* in five Ethiopian populations and a 4448 base pair region re-sequenced in eight additional sub-Saharan African populations. Tables 2.2 and 4.1 have information on populations analysed in this chapter. All PCR and re-sequencing conditions are as outlined in section 2.2. Within the re-sequenced regions, all identified promoter, exonic, splice site and 3´ UTR variants were analysed for potential effects on *CYP3A5* transcription and protein expression, using bioinformatics software (see section 2.3.9). For all analyses, the ancestral allele at a given nucleotide position was inferred from the NCBI chimpanzee *CYP3A5* reference sequence.

5.1.1 Previously reported variation at the CYP3A5 locus

To date a total of 258 variants have been identified within *CYP3A5* (from the 5´ UTR of exon one to the 3´ UTR) and 114 of these occur at a global frequency of ≥1% (http://www.ncbi.nlm.nih.gov/). A further 9 variants have been identified within 2500 base pairs upstream of the ATG start codon: 4 occur in the proximal promoter; none of which occur in experimentally established transcription factor binding sites and so are unlikely to affect the initiation of *CYP3A5* transcription (Nakamura et al. 2001; Xie et al. 2004). 15 variants have been identified within a 2385 base pair region immediately downstream of the 3´ UTR of *CYP3A5.* Of the common variants (i.e. with a global frequency of ≥1%), the majority occur in non-coding regions. Figure 5.1 shows the distribution of all identified variants across the

CYP3A5 gene region; Table 5.1 provides details about the type of polymorphisms identified in the *CYP3A5* gene and information about the size of each genic region, as listed on the NCBI database (http://www.ncbi.nlm.nih.gov/). Figure 5.2 shows the outline of the *CYP3A5* gene with proportions of common variation (≥1% global frequency) annotated.

Figure 5.1: The distribution of all identified variants across the *CYP3A5* gene region. Data have been compiled from the NCBI database (http://www.ncbi.nlm.nih.gov/).



A comparison of the coding regions (ATG start codon to the 3´ UTR) of the four *CYP3A* genes: *CYP3A4, CYP3A5, CYP3A7* and *CYP3A43,* found that *CYP3A4* has the highest number of identified variants (284) compared to its paralogues *CYP3A5* (218), *CYP3A7* (181) and *CYP3A43* (273). Although only 28% of all identified *CYP3A4* variants are observed at a global frequency of ≥1%; in contrast to 43% for *CYP3A5*, 65% for *CYP3A7* and 82% for *CYP3A43* (http://www.ncbi.nlm.nih.gov/). Chi-squared tests, with Yates' correction, found that *CYP3A4* has significantly more polymorphic sites, relative to fixed sites, than the remaining *CYP3A* genes (Table 5.2). Fisher's exact tests were performed to examine whether there are significant inter-*CYP3A* differences in the ratios of common variation (occurs at a global frequency of ≥1%) to rare variants (presented in Table 5.3). *CYP3A4* and *CYP3A5* both have significantly fewer common variants relative to *CYP3A7* and *CYP3A43*.

Table 5.1: A summary of global variation at the *CYP3A5* locus, plus 2500 base pairs either side, as reported by NCBI (http://www.ncbi.nlm.nih.gov/)

CYP3A5 region	Size of region (base pairs)	Total number of variants reported	Substitutions	Insertions/ Deletions	Microsatellite repeats	Known regulatory motifs	Consensus Splice sites	Synonymous	Non-synonymous	Frameshift
5´ upstream	1700	11	8	3	0	-	-	-	-	-
Proximal promoter	800	11	11	0	0	0	-	-	-	-
3´ untranslated region	114	3	3	0	0	0	0	-	-	0
3´ downstream	2386	15	12	2	1	-	-	-	-	-
Exon 1	71	0	-	-	-	-	-	-	-	-
Exon 2	94	5	4	1	0	-	-	4	0	1
Exon 3	53	0	-	-	-	-	-	-	-	-
Exon 4	100	2	2	0	0	-	-	2	0	0
Exon 5	114	0	-	-	-	-	-	-	-	-
Exon 6	89	0	-	-	-	-	-	-	-	-
Exon 7	149	3	3	0	0	-	-	2*	1	0
Exon 8	128	0	-	-	-	-	-	-	-	-
Exon 9	67	2	2	0	0	-	-	1	1	0
Exon 10	161	1	1	0	0	-	-	0	1	0
Exon 11	227	4	2	2	0	-	-	0	2	2**
Exon 12	160	2	2	0	0	-	-	0	2	0
Exon 13	96	2	2	0	0	-	-	0	2	0
Intron 1	3617	32	26	6	0	-	0	-	-	-
Intron 2	1529	11	10	1	0	-	0	-	-	-
Intron 3	1853	15	14***	1	0	-	0	-	-	-

Intron 4	5514	39	31	7	1	-	0	-	-	-
Intron 5	262	5	5	0	0	-	0	-	-	-
Intron 6	1286	7	6	1	0	-	0	-	-	-
Intron 7	1070	7	7	0	0	-	0	-	-	-
Intron 8	1085	9	9	0	0	-	0	-	-	-
Intron 9	2156	14	14	0	0	-	0	-	-	-
Intron 10	7719	31	23	7	1	-	0	-	-	-
Intron 11	2320	12	12	0	0	-	0	-	-	-
Intron 12	1672	15	12	3	0	-	0	-	-	-

Note:

* A synonymous substitutions in exon 7 defines the *CYP3A5*6* variant, which has been reported to be associated with low/non-expression of CYP3A5 protein by causing skipping of exon 7 (Hustert et al. 2001; Kuehl et al. 2001).

** A frameshift mutation in exon 11 defines the *CYP3A5*7* variant, which has been reported to be associated with low/non-expression of CYP3A5 due to the creation of a premature termination codon (Hustert et al. 2001; Xie et al. 2004)

*** A substitution in intron 3 defines the *CYP3A5*3* variant, which has been experimentally shown to reduce *in vivo* CYP3A5 expression, in some cases to undetectable levels (Lin et al. 2002; Busi and Cresteil 2005)

Figure 5.2: A diagram of the distribution of variants reported in the *CYP3A5* gene.

- The pink boxes represent *CYP3A5* exons, although they are not spaced according to scale. The red box represents the proximal promoter region of *CYP3A5* and the green box is the 3´ untranslated region (UTR). Spacer regions, represented by black lines between adjacent exons, are introns.
- X*n* refers to a specific exon of number *n.*
- The 5´ upstream and 3´ downstream regions either side of the gene correspond to 1700 base pairs and 2386 base pairs, respectively.
- The exons are ordered from Exon 1-13, although they are not all numbered on the Figure.
- The numbers in the brackets correspond to the total number of variants reported, on the NCBI database (http://www.ncbi.nlm.nih.gov/), in each region of *CYP3A5*. The values preceding the "/" indicate the number of variants which occur at a global frequency of ≥1%, the numbers after the "/" sign correspond to the total number of reported variants in each *CYP3A5* region.

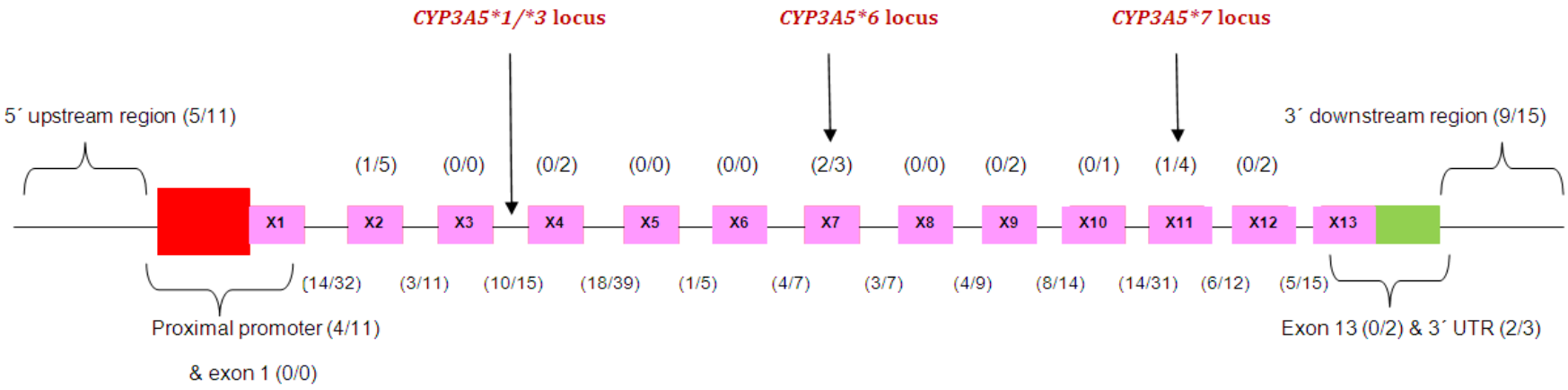


Table 5.2: The results of chi-squared tests which compared pairwise differences in the *CYP3A* ratios of polymorphic to fixed sites. Yates' correction was applied for each comparison. The data compared have been compiled from the NCBI database, build 132 (http://www.ncbi.nlm.nih.gov/). Statistically significant differences at the 5% level are highlighted in green and shown in bold.

	CYP3A4	*CYP3A5*	*CYP3A7*	*CYP3A43*
CYP3A4	*			
CYP3A5	**<0.0001**	*		
CYP3A7	**0.0001**	0.2516	*	
CYP3A43	**0.0001**	0.6675	0.0970	*

Table 5.3: The results of pairwise Fisher's exact comparisons of the ratios of common variation (observed at a global frequency of ≥1%) to rare variation in *CYP3A* genes. The data compared have been compiled from the NCBI database, build 132 (http://www.ncbi.nlm.nih.gov/). Statistically significant differences at the 5% level are highlighted in green and shown in bold.

	CYP3A4	*CYP3A5*	*CYP3A7*	*CYP3A43*
CYP3A4	*			
CYP3A5	**0.0135**	*		
CYP3A7	**<0.0001**	**0.0173**	*	
CYP3A43	**<0.0001**	**<0.0001**	0.1211	*

Nine *CYP3A5* variants, which include *CYP3A5*3, CYP3A5*6* and *CYP3A5*7*, have been reported to be candidates for low/non enzyme expression (Xie et al. 2004). To date, *in vivo* studies have established a role for *CYP3A5*3* in affecting enzyme activity and function (Givens et al. 2003); this is due to high frequencies of this variant observed in many populations, and a bias towards genotyping of this variant alone in previous medical association studies. There are fewer data for the effect of *CYP3A5*6* and *CYP3A5*7* on enzyme activity and expression, although CYP3A5 catalytic activity, measured by Western blot, was observed to be lower in *CYP3A5*6* heterozygotes than in *CYP3A5*1* homozygotes, see section 1.2.3.1 (Kuehl et al. 2001). *CYP3A5*7* was also observed to be associated with low/non hepatic expression of CYP3A5 (Hustert et al. 2001). A previous study identified an anomalous *CYP3A5*1/*3* heterozygote with undetectable levels of CYP3A5 protein. The individual had a *CYP3A5*7* mutation on the *CYP3A5*1* chromosome; providing further evidence that this allele is associated with decreased CYP3A5 expression (Givens et al. 2003).

For the remaining six candidate low/non expression alleles, *in vitro* data are available. However due to the lower global frequencies of these alleles, comparative to *CYP3A5*3, CYP3A5*6* and *CYP3A5*7*, it has been difficult to elucidate the independent effects of these alleles *in vivo* (Xie et al. 2004) and they are often observed in populations with high frequencies of *CYP3A5*3* (Table 5.4).

There is a general problem with examining the effect of *CYP3A5* alleles on hepatic and intestinal drug clearance. There is considerable substrate overlap between CYP3A5, CYP3A4 and a membrane efflux transporter protein P-glycoprotein, present in many tissues in which CYP3A enzymes act, encoded by the *MDR1* gene (Kim et al. 1999; Kim et al. 2008). This makes it difficult to examine the *in vivo* effects of CYP3A5 protein alone; *in vitro* experiments can elucidate the effects of specific variants, on protein activity, but may not reflect what actually happens *in vivo*. There are some examples of where there is a strong association between CYP3A5 activity and disease pathology, such as salt-sensitive hypertension (Givens et al. 2003; Fromm et al. 2005; Bochud et al. 2006; Bochud et al. 2009), and adverse clinical outcomes associated with drug therapy, such as with the immunosuppressant drug tacrolimus (Zhao et al. 2005), although the metabolism of tacrolimus is also influenced by the *MDR1* gene and P-glycoprotein (Zheng et al. 2003).

5.1.2 Characteristics of variation at the CYP3A5 locus

Linkage disequilibrium (LD) across *CYP3A5* varies between populations genotyped as part of the HapMap consortium (http://www.hapmap.org/), see Figure 5.3. LD is high across Tuscans from Italy; individuals of Mexican ancestry from Los Angeles; Japanese individuals from Tokyo; Gujarati Indians from Houston Texas; Chinese individuals from Metropolitan Denver and from Beijing; and individuals with Northern and Western European ancestry in the United States; all non-African populations. This is consistent with previous studies reporting that LD is higher in populations outside of Africa (Conrad et al. 2006).

There is strong LD across the ~220kb *CYP3A* locus, see Figure 5.4. One of the most comprehensive studies of human variation at the *CYP3A* locus found that the *CYP3A5*3* mutation defined 84% of the *CYP3A* haplotypes observed in Han Chinese individuals and 98% of those in European individuals (Thompson et al. 2006). Additionally in African-Americans the authors observed that a number of *CYP3A* haplotypes were defined by high LD between the *CYP3A5*1* allele and a *CYP3A7*2* variant, which occurs in exon 11 of the *CYP3A7* gene and causes increased enzyme expression (Rodriguez-Antona et al. 2005) and it was suggested that tight linkage between these two variants, and increased expression of both CYP3A5 and CYP3A7 enzymes, may provide a selective advantage in the foetus.

Table 5.4: A list of *CYP3A5* alleles that have been reported to be candidates for causing low/non protein expression; "n" refers to the number of individuals sampled.

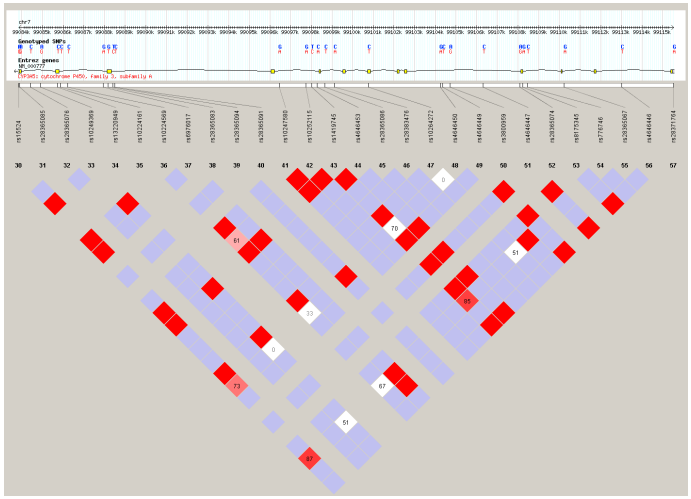
Variant name	Variant	Region of gene	Coding/non-coding	Evidence for low/non expression	Population and variant allele frequency	Reference
*CYP3A5*2*	C>A	Exon 11	Coding: non-synonymous; T398N	*In vitro* observation found 2/5 livers not expressing CYP3A5 protein were homozygous for *CYP3A5*2,* although it is unknown whether this allele is tightly linked to *CYP3A5*3* due to its low global frequency	French Caucasians (1.25%; *n*=5)	(Jounaidi et al. 1996; Chou et al. 2001; Lee et al. 2003)
*CYP3A5*3*	A>G	Intron 3	Non-coding, reported to cause part of intron 3 to be retained in the mature mRNA	*In vitro* and *in vivo* data, including medical association studies with the immunosuppressant drug tacrolimus	See Table 3.8	(Hustert et al. 2001; Kuehl et al. 2001; Busi and Cresteil 2005)
*CYP3A5*4*	A>G	Exon 7	Coding: non-synonymous; Q200R	No functional *in vitro* data due to the low frequency of the variant	Chinese (0.36% *n*=110)	(Chou et al. 2001)
*CYP3A5*5*	T>C	Intron 5	This transition occurs in intron 5 immediately adjacent to the intron 5 donor splice site.	An *in vitro* study found that *CYP3A5*5* produced multiple splicing products in Caco-2 cell lines	Chinese (0.99%; *n*=110)	(Chou et al. 2001; Lamba et al. 2002)
*CYP3A5*6*	G>A	Exon 7	Coding: synonymous; reported to create a splice site which causes skipping of exon 7	cDNA extracted from *CYP3A5*1/CYP3A5*6* hepatocytes did not have the sequence for exon 7. Heterozygotes also had lower catalytic activity towards midazolam (a known CYP3A substrate) than *CYP3A5*1* homozygotes	See Table 3.8	(Hustert et al. 2001; Kuehl et al. 2001)
*CYP3A5*7*	T insertion	Exon 11	Coding: non-synonymous change; creates premature termination codon at position 348	*CYP3A5*1/CYP3A5*7* individuals have lower catalytic activity towards midazolam than *CYP3A5*1* homozygotes	See Table 3.8	(Hustert et al. 2001)
*CYP3A5*8*	C>T	Exon 2	Coding: non-synonymous change; R28C	*In vitro* studies found that CYP3A5 with the *CYP3A5*8* mutation, purified from an *Escherichia coli* system, had decreased ability to metabolise testosterone (a known CYP3A substrate)	Africans (4%; *n*=24)	(Lee et al. 2003)
*CYP3A5*9*	G>A	Exon 10	Coding: non-synonymous change; A337T	*In vitro* studies found that CYP3A5 with the *CYP3A5*9* mutation, purified from an *Escherichia coli* system, had decreased ability to metabolise testosterone (a known CYP3A substrate)	East Asians (2%; *n*=24)	(Lee et al. 2003)
*CYP3A5*10*	T>C	Exon 12	Coding: non-synonymous change; F446S	*In vitro* studies found that CYP3A5 with the *CYP3A5*10* has almost complete loss of activity. This is because the mutation occurs within the heme binding site of CYP3A5, which is essential for P450 mediated drug metabolism.	Caucasians (2%; *n*=24)	(Lee et al. 2003)

N.B. The 24 Africans genotyped in the 2003 paper by Lee et al (Lee et al. 2003) comprise 15 African-Americans and 9 African Pygmies; the 24 Asians are 4 Indo-Pakistani, 5 native Taiwanese, 5 mainland Chinese from Beijing, 3 Cambodians, 3 Japanese and 4 Melanesian; and the 24 Caucasians are from the USA and Europe.

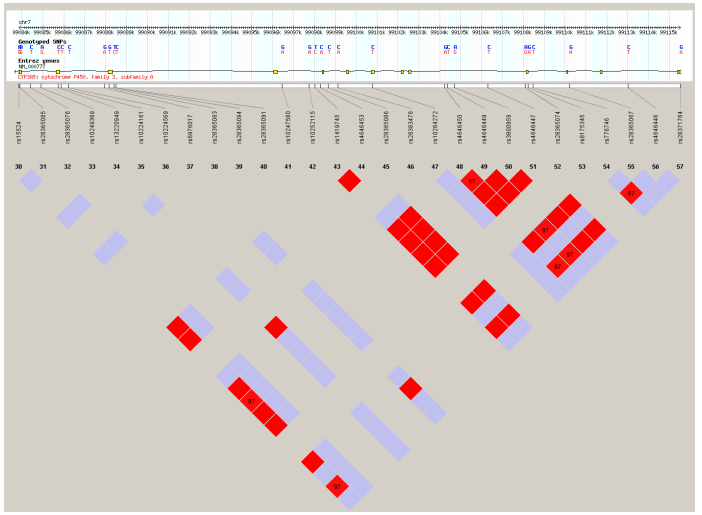
Interestingly, across the *CYP3A* locus, nucleotide diversity per base pair is much lower for *CYP3A5,* than for the remaining genes, in every re-sequenced population (Table 5.5). Additionally comparisons of all *CYP3A* genes found that *CYP3A5* is characterised by an excess of rare variants in each of the African-American, European and Han Chinese populations; and had the lowest amount of sequence divergence from the chimpanzee (Thompson et al. 2006). LD is highest across the *CYP3A* cluster in populations in which the *CYP3A5*3* allele is observed at high frequencies. These populations also have lower measures of nucleotide diversity comparative to those in which the *CYP3A5*1* allele is at higher frequencies.

Genome scans have identified candidates for positive selection on human chromosome 7 (Sabeti et al. 2006). An excess of rare variants and regions of extended haplotype homogeneity have been identified in a region surrounding the Kell blood antigen cluster (Sabeti et al. 2006) and a functional polymorphism in the gene encoding the antigen *CD36;* found at high frequency in African populations and associated with differential susceptibility to cerebral malaria (Aitman et al. 2000; Pain et al. 2001). *CYP3A5* has also been proposed as a target of selection (Thompson et al. 2004; Sabeti et al. 2006); given the strong LD observed within the region and the excess of rare variants at the *CYP3A5* locus in these populations (Thompson et al. 2006). However it is important to note that LD across chromosome 7 may extend beyond the *CYP3A* cluster and there may be an alternative target of selection, tightly linked to the *CYP3A5*3* allele, which has driven it to high frequencies in certain populations. Additionally, it is possible that *CYP3A5*3* is a young allele which has recently evolved on an existing haplotype background with a paucity of variation; and may mimic a signal of selection. The age of *CYP3A5*3* is addressed in chapter 7.

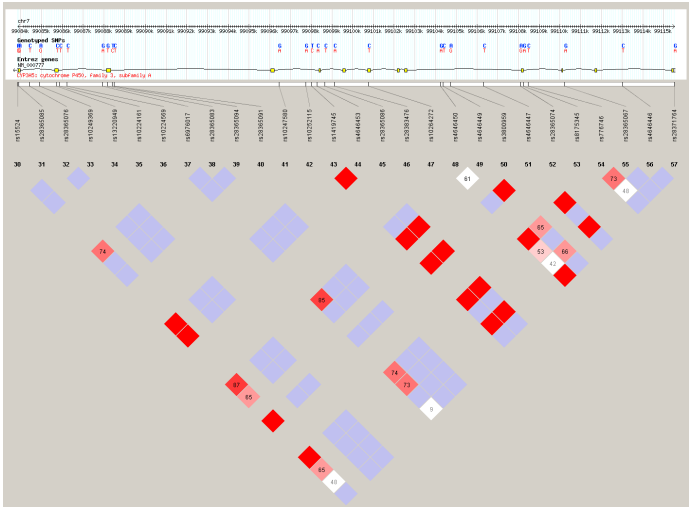
Figure 5.3: Linkage disequilibrium at the *CYP3A5* locus in 11 populations, genotyped as part of the HapMap consortium (http://www.hapmap.org/). Values in boxes refer to *D´;* where there is no number the value of *D´* is equal to 1. Red and pink boxes indicate a significant *D´* value; blue and white indicate non-significant.



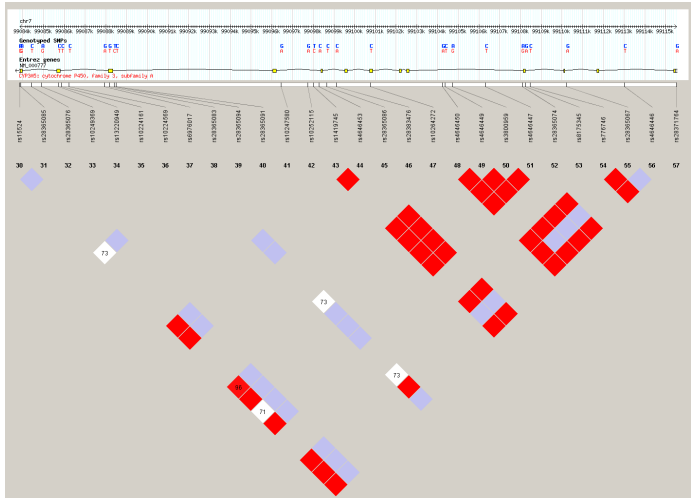
African ancestry in Southwest USA



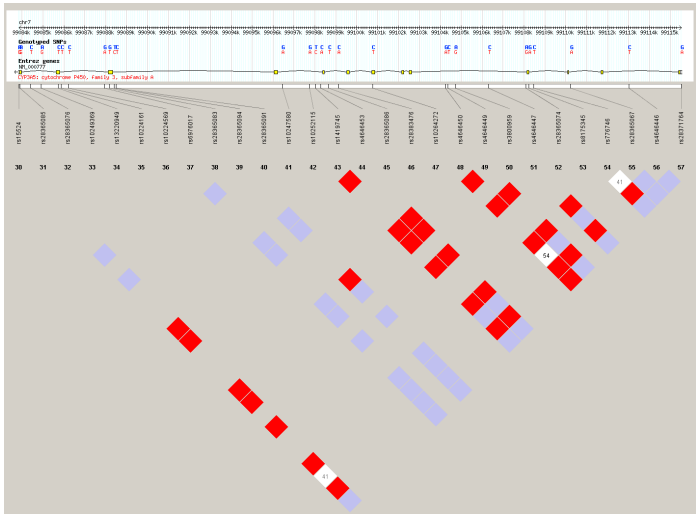
Han Chinese in Beijing, China



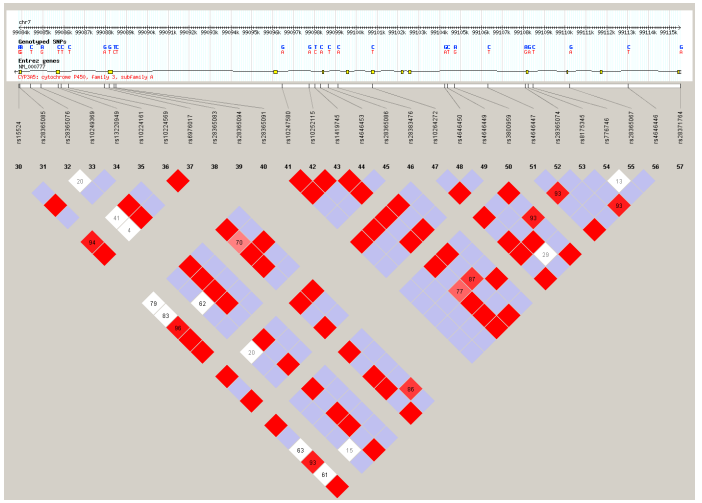
Utah residents with Northern and Western European ancestry from the CEPH collection



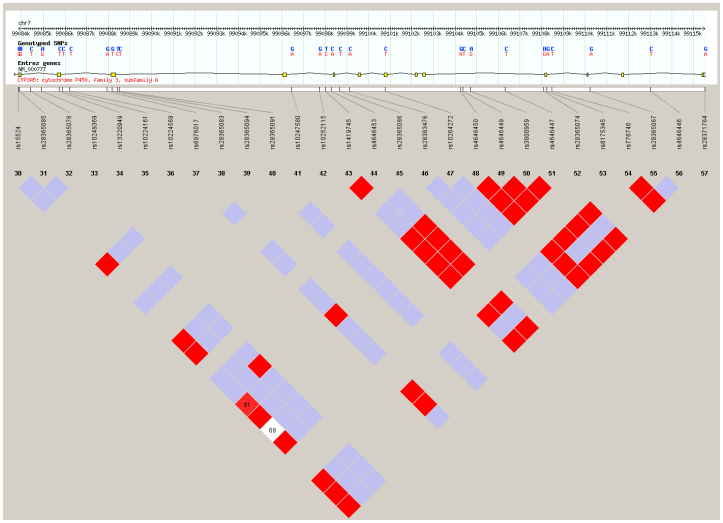
Chinese in Metropolitan Denver, Colorado



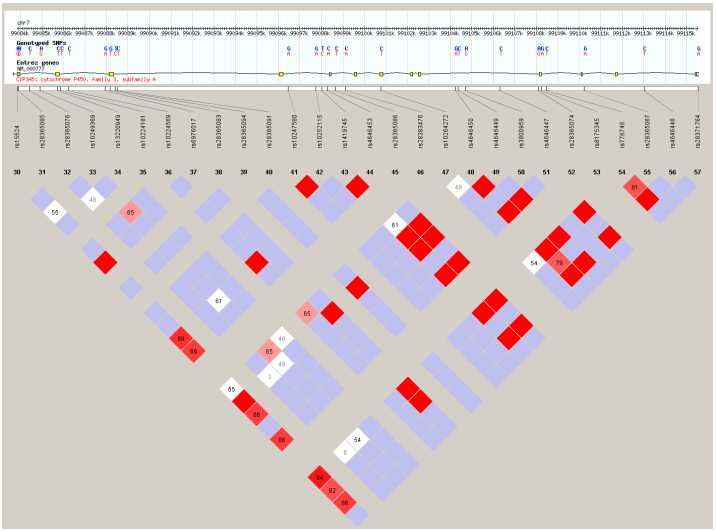
Gujarati Indians in Houston, Texas



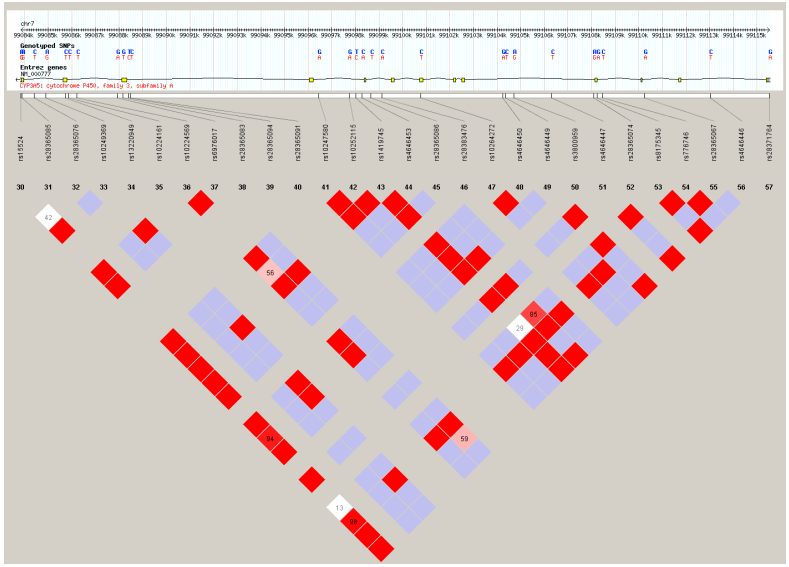
Luhya in Webuye, Kenya



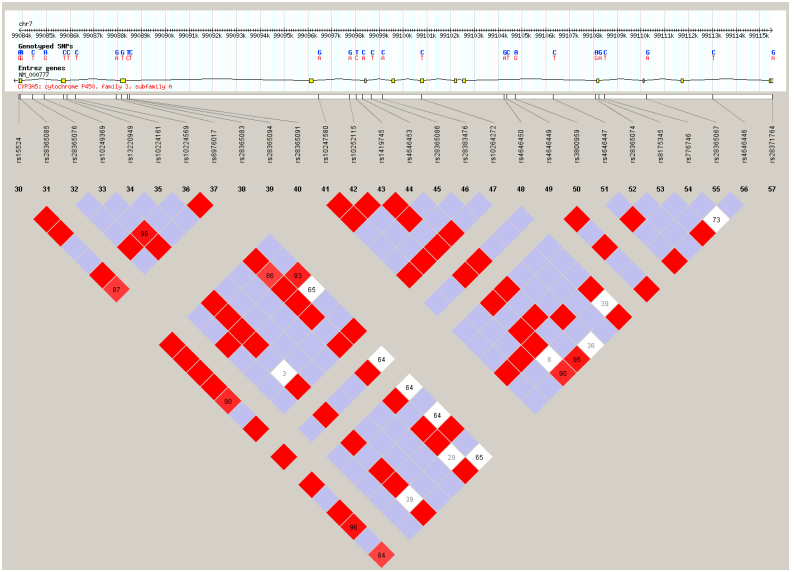
Japanese in Tokyo, Japan



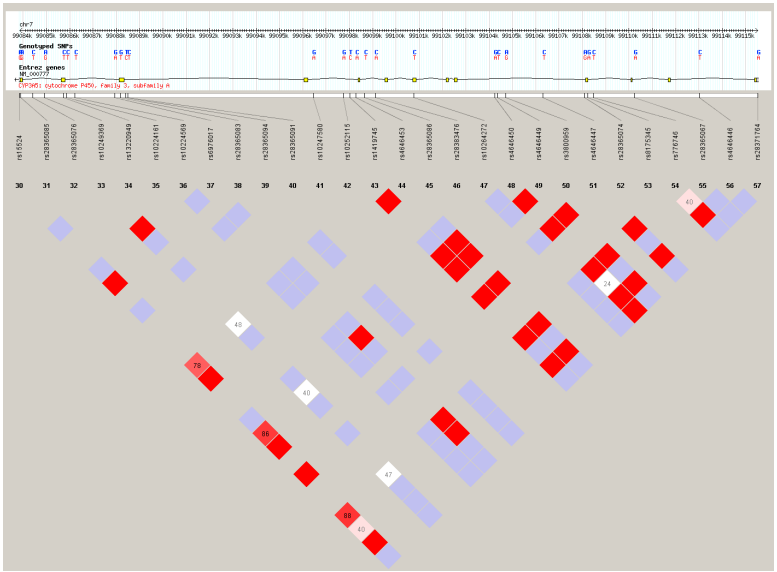
Mexican ancestry in Los Angeles, California



Maasai in Kinyawa, Kenya



Yoruba in Ibadan, Nigeria



Tuscans in Italy

Figure 5.4: The extent of linkage disequilibrium across the entire *CYP3A* cluster of genes in three ethnically distinct global populations. The image has been taken from (Thompson et al. 2006). Each row is a haplotype and each column is a polymorphic site. Rows with grey shading are the corresponding chimpanzee sequence, which was used to infer the ancestral allele at each polymorphic position. Derived variants are show in blue and ancestral in yellow. Note the lack of variation in the *CYP3A5* gene (far right hand side). The most common variant observed, in each of the three populations, is the *CYP3A5*3* allele; which defined over 80% of all *CYP3A* haplotypes in the cohort.

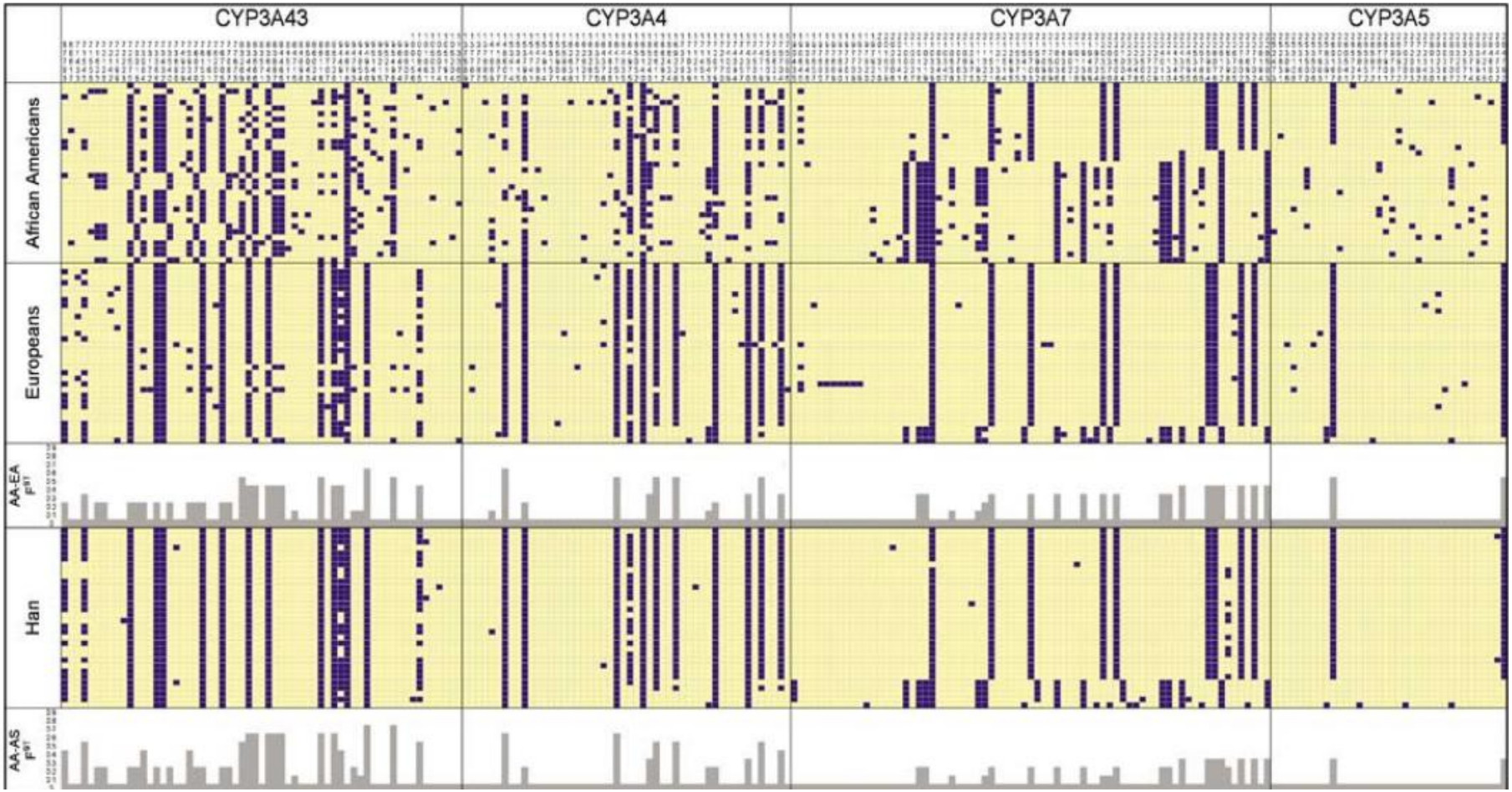


Table 5.5: Summary statistics of polymorphism data from African-American, Han Chinese and European individuals re-sequenced at the entire *CYP3A* cluster. The image has been taken from (Thompson et al. 2006). **Note** the low measures of nucleotide diversity for *CYP3A5,* in each of the three populations, comparative to the other *CYP3A* genes, and the negative values of Tajima's *D* (annotated on the Figure).

Table 1 Summary statistics of polymorphism data from African-American, European, and Han samples

Gene	*Length (bp)*	*Summary statistics by population sample*													
		African Americans[a]					*Europeans*[a]					*Han*[a]			
		S^b	π^c	D^d	$\theta_W{}^e$	θ_W:*div*[f]	S^b	π^c	D^d	$\theta_W{}^e$	θ_W:*div*[f]	S^b	π^c	D^d	$\theta_W{}^e$
CYP3A43	17 825	49	0.78	0.51 (96)	0.68	7% (8)	28	0.29	−0.87 (10)	0.39	4% (12)	9	0.12	−0.04	0.13
CYP3A4	16 822	36	0.48	−0.33 (66)	0.53	6% (3)	22	0.16	−1.82 (3)	0.32	4% (10)	9	0.09	−0.99	0.13
CYP3A7	19 071	44	0.72	0.97 (99)	0.57	6% (3)	40	0.28	−1.65 (3)	0.52	5% (27)	37	0.38	−0.77	0.48
CYP3A5	14 582	22	0.25	−1.14 (14)	0.37	5% (1)	13	0.08	−2.15 (1)	0.22	3% (3)	7	0.06	−1.37	0.11

[a] Samples are a subset of those used in a previous study of *CYP3A4* and *CYP3A5*.[20] Here, the summary statistics for these two genes and their percentile rank relative to the Seattle SNP genes were re-calculated for the sample subset used in this analysis.
[b] Number of segregating sites.
[c] Nucleotide diversity per bp ($\times 10^{-3}$).
[d] Tajima's *D*.[30] Numbers in parentheses indicate the percentile rank of the *D* value relative to the distribution of *D* values in the Seattle SNP genes.
[e] Watterson's estimator of the population mutation rate parameter θ ($=4N\mu$) per bp ($\times 10^{-3}$).[29]
[f] Ratio of θ_W to the amount of sequence divergence between human and chimpanzee. Numbers in parentheses indicate the percentile rank of the θ_W:div value relative to the distribution of θ_W:div values in the Seattle SNP genes.

5.2 Variation at the *CYP3A5* locus in Ethiopia

5.2.1 Variation across the entire CYP3A5 locus

76 polymorphic sites were identified within a 12,237 base pair region of chromosome 7, which includes the *CYP3A5* promoter, coding region and 3′ untranslated region (UTR), re-sequenced in five Ethiopian populations. A diagram showing the distribution of all polymorphic sites observed in the 12,237 base pair region is presented in Figure 5.5. 50 of all observed variants (~65.79%) occur in introns 3 and 9, the *CYP3A5* promoter and in ≥1700 base pairs of sequence either side of the gene. Table 5.6 shows the full list of polymorphic sites identified in the five Ethiopian groups and indicates which are novel and which have been previously identified. 69 variants were single nucleotide substitutions; 5 were deletions (2 single nucleotide deletions, 1 double nucleotide deletion and 2 triple nucleotide deletions); 1 insertion (a single base in exon 11); and 1 microsatellite repeat were also observed. A total of 13 variants, excluding singletons, were private to specific Ethiopian populations.

51 polymorphic sites were identified within an 8571 base pair region comprising the *CYP3A5* promoter, full coding region and 3′ UTR, and 16 of these were singletons. Interestingly, every observed exonic polymorphism was a singleton, and the highest frequency variants were *CYP3A5*3* and *CYP3A5*6*. Excluding the singletons, 8 variants were private to specific Ethiopian groups.

The majority (53%) of all identified variants were observed at a frequency of <0.66% of chromosomes (n=758), see Figure 5.5 and 5.6. 23 of the 76 variants (~30.3%) are singletons, i.e. they were observed in a single heterozygous individual. The only variants identified on over 50% of all Ethiopian chromosomes were the allele defining *CYP3A5*3* and rs15524.

Figure 5.5: A full list of all polymorphic sites identified across the 12,237 base pair *CYP3A5* region re-sequenced in five Ethiopian populations.

- The pink boxes represent *CYP3A5* exons, although they are not spaced according to scale. The red box represents the proximal promoter region of *CYP3A5* and the green box is the 3´ untranslated region (UTR). Spacer regions, represented by black lines between adjacent exons, are introns.
- The 5´ upstream and 3´ downstream regions either side of the gene correspond to 1700 base pairs and 2385 base pairs, respectively.
- The exons are ordered from Exon 1-13, although they are not all numbered on the Figure.
- The numbers in the brackets correspond to the total number of variants observed in five Ethiopian populations.

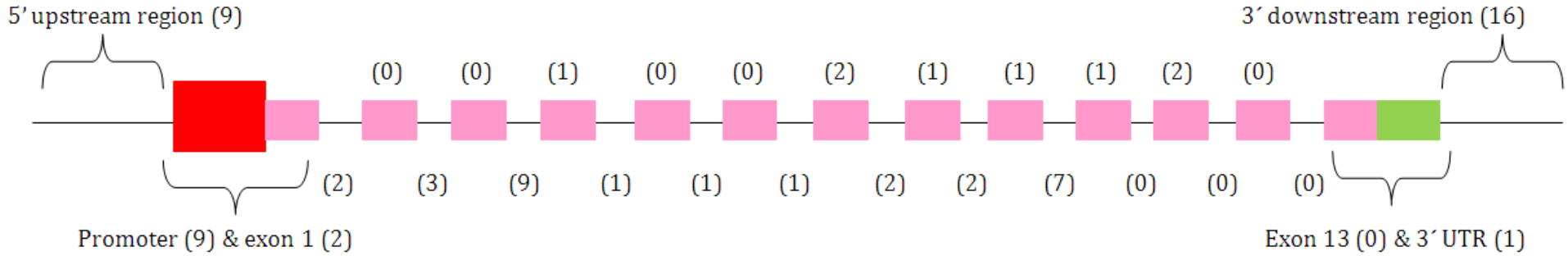


Table 5.6: A list of all polymorphic sites identified in a 12,237 base pair *CYP3A5* region re-sequenced in five Ethiopian populations.

- For each Ethiopian population: *n* refers to the total number of chromosomes on which a particular variant was observed and *f* is the relative frequency a variant was observed in a particular Ethiopian group and was calculated by dividing *n* for a population by *n* for the Ethiopian cohort.
- "Total re-sequenced" refers to the total number of chromosomes, in the Ethiopian cohort, successfully re-sequenced at a particular site. In the "Total" columns; *n* refers to the total number of chromosomes on which a variant allele(s) was seen and *f* is the frequency within the entire Ethiopian cohort.
- light blue shading indicates novel mutations

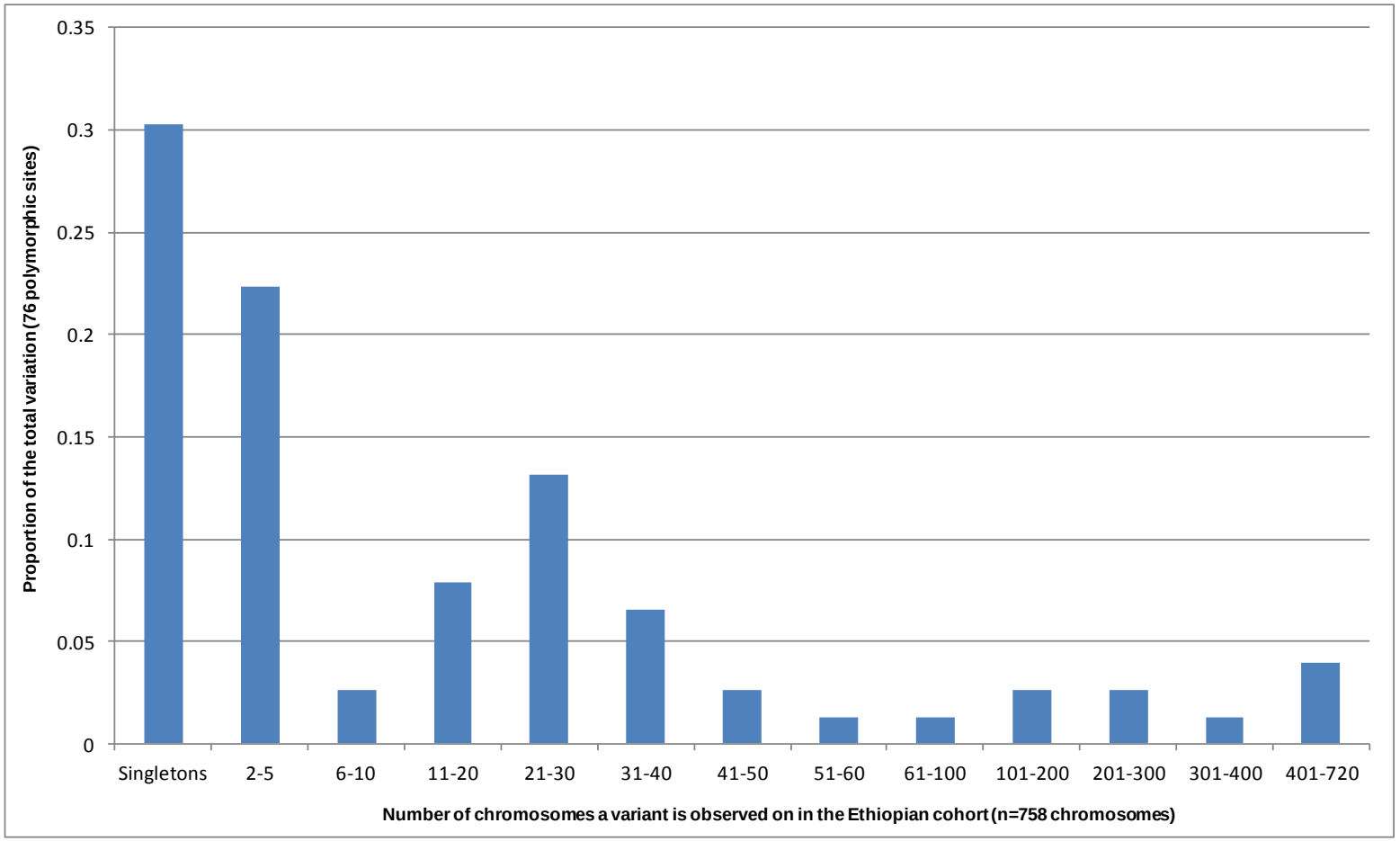
					Afar		Amhara		Anuak		Maale		Oromo		Total		
Region of *CYP3A5*	Position on chromosome 7: (NCBI Build 132, February 2009)	*CYP3A5* variant and its position relative to the translation initiation codon (A of ATG is +1)	NCBI dbSNP database refSNP ID	Effect	*f*	n	*f*	n	*f*	n	*f*	n	*f*	n	*f*	n	Total re-sequenced
5' upstream region	99279710	-2191 C>T	rs10270499		0.04	28	0.03	22	0.05	37	0.03	23	0.03	19	0.1702	129	758
5' upstream region	99279307	-1788 G>A			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0027	2	754
5' upstream region	99279136	-1617 T>C	rs776741		0.05	38	0.05	38	0.10	75	0.07	53	0.05	38	0.3210	242	754
5' upstream region	99279051	-1532 T>C			0.00	0	0.00	0	0.00	0	0.01	4	0.00	0	0.0053	4	754
5' upstream region	99278876	-1357 C>T			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	758
5' upstream region	99278862	-1343 T>A			0.00	1	0.00	3	0.01	10	0.00	3	0.01	6	0.0303	23	758
5' upstream region	99278827	-1308 C>T			0.00	1	0.00	3	0.01	10	0.00	3	0.01	6	0.0303	23	758
5' upstream region	99278771	-1252 1 base pair deletion			0.00	0	0.00	2	0.00	0	0.00	0	0.00	0	0.0026	2	758
5' upstream region	99278522	-1003 A>C	rs36231118		0.01	4	0.00	3	0.02	15	0.01	9	0.01	4	0.0462	35	758
Proximal promoter	99278314	-795 T>A	rs3823812		0.00	3	0.00	3	0.01	4	0.01	10	0.01	5	0.0331	25	756
Proximal promoter	99278267	-748 C>G			0.01	5	0.00	2	0.00	1	0.00	1	0.01	6	0.0198	15	756
Proximal promoter	99278224	-705 3 base pair deletion			0.00	1	0.00	1	0.01	5	0.00	1	0.00	3	0.0146	11	756
Proximal promoter	99278223	-704 A>G			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0013	1	756

Proximal promoter	99278152	-633 C>A			0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.0013	1	756
Proximal promoter	99278146	-627 G>A			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	756
Proximal promoter	99278144	-625 A>G			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0013	1	756
Proximal promoter	99278070	-551 C>A	rs28365079		0.01	4	0.01	5	0.02	15	0.01	8	0.01	4	0.0476	36	756
Proximal promoter	99277988	-469 G>A			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0013	1	756
5' UTR of exon 1	99277593	-74 C>T	rs28371764		0.00	2	0.01	6	0.00	0	0.00	2	0.00	2	0.0158	12	758
5' UTR of exon 1	99277544	-25 A>C			0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.0013	1	758
Intron 1	99277392	127 G>A			0.00	0	0.00	0	0.00	1	0.00	2	0.00	0	0.0040	3	758
Intron 1	99277337	182 C>A			0.00	0	0.00	0	0.00	3	0.00	0	0.00	0	0.0040	3	758
Intron 2	99272310	5209 C>T	rs28365067		0.01	11	0.02	12	0.01	5	0.01	8	0.01	8	0.0580	44	758
Intron 2	99272290	5229 G>A	rs41301652		0.00	0	0.00	0	0.00	2	0.00	0	0.00	0	0.0026	2	758
Intron 2	99272275	5244 C>T			0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.0026	2	758
Intron 3	99272103	5416 C>T			0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.0026	2	758
Intron 3	99272009	5510 T>A	rs28969392		0.01	6	0.01	4	0.01	10	0.01	9	0.00	3	0.0422	32	758
Intron 3	99271928	5591 C>T	rs41301655		0.00	0	0.01	4	0.00	1	0.00	0	0.00	2	0.0092	7	758
Intron 3	99271853	5666 A>G	rs41301658		0.00	1	0.00	1	0.00	3	0.01	7	0.00	2	0.0185	14	758
Intron 3	99271808	5711 A>G	rs41258334		0.01	11	0.01	11	0.01	5	0.01	9	0.01	8	0.0580	44	758
Intron 3	99271778	5741 A>G			0.01	6	0.00	3	0.01	4	0.01	8	0.00	3	0.0317	24	758
Intron 3	99270539	6980 A>G	rs776746	Defines the allelic variant *CYP3A5*3*	0.13	95	0.14	102	0.06	44	0.10	75	0.13	97	0.5581	413	740
Intron 3	99270504	7015 3 base pair deletion			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0014	1	740
Intron 3	99270318	7201 C>T	rs8175345		0.00	0	0.00	1	0.01	9	0.00	0	0.00	1	0.0149	11	740
Exon 4	99270249	7270 G>A		G77S	0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0014	1	740
Intron 4	99270164	7355 C>T	rs28365074		0.00	0	0.00	0	0.00	1	0.00	0	0.00	2	0.0041	3	740
Intron 5	99264352	13167 T>C	rs68178885		0.00	3	0.00	2	0.00	1	0.00	3	0.00	1	0.0132	10	758
Intron 6	99264149	13370 G>A	rs41301670		0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.0027	2	754

Exon 7	99262835	14684 G>A	rs10264272	Defines the allelic variant *CYP3A5*6*	0.04	28	0.03	23	0.05	39	0.03	23	0.03	21	0.1763	134	760
Exon 7	99262793	14726 A>G	rs2838372	Synonymous	0.00	1	0.00	0	0.00	0	0.00	0	0.00	0	0.0013	1	760
Intron 7	99262642	14877 A>G			0.00	1	0.01	5	0.02	12	0.01	9	0.00	2	0.0382	29	760
Intron 7	99261737	15782 T>C	rs28969393		0.01	5	0.01	4	0.01	9	0.01	9	0.00	3	0.0396	30	758
Exon 8	99261651	15868 A>G		K266R	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	758
Intron 8	99261583	15936 C>A			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0026	2	758
Intron 8	99260546	16973 G>A			0.00	0	0.00	1	0.00	0	0.00	0	0.00	0	0.0013	1	756
Exon 9	99260502	17017 C>T		R268Stop	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	756
Intron 9	99260407	17112 C>T	rs28383478		0.00	0	0.00	2	0.00	0	0.00	0	0.00	0	0.0026	2	756
Intron 9	99260362	17157 G>T	rs4646453		0.00	3	0.00	3	0.01	4	0.01	10	0.01	5	0.0331	25	756
Intron 9	99260282	17237 T>G			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	756
Intron 9	99260170	17349 T>G			0.00	3	0.00	2	0.01	7	0.01	7	0.00	3	0.0291	22	756
Intron 9	99258524	18995 C>T	rs10247580		0.00	0	0.00	2	0.02	12	0.01	7	0.00	1	0.0291	22	756
Intron 9	99258320	19199 G>A			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	756
Intron 9	99258316	19203 T>C			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0026	2	756
Exon 10	99258124	19395 A>C		K342T	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.0013	1	756
Exon 11	99250397	27125-27126 T insertion	rs41303343	Defines the allelic variant *CYP3A5*7*	0.00	0	0.00	0	0.00	1	0.00	1	0.00	0	0.0026	2	760
Exon 11	99250381	27138 A>G		V350M	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	760
Intron 12	99247647	29872 G>T			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0026	2	756
Intron 12	99247503	30016 1 base pair deletion	rs28365093		0.00	3	0.01	4	0.02	15	0.01	8	0.01	4	0.0450	34	756
Intron 12	99246026	31493 T>C	rs28365069		0.01	4	0.01	11	0.01	11	0.02	18	0.01	9	0.0699	53	758
3' UTR	99245914	31605 C>T	rs15524	None, although is often on the background of rs776746	0.14	105	0.14	109	0.09	69	0.11	84	0.14	107	0.6253	474	758
3' downstream region	99245796	31723 T>A			0.00	1	0.00	0	0.00	0	0.00	0	0.00	0	0.0013	1	758

3' downstream region	99245707	31812 G>A		0.00	0	0.00	1	0.00	0	0.00	0	0.00	0	0.0013	1	758
3' downstream region	99245537	31982 2 base pair deletion		0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	758
3' downstream region	99245499	32020 C>T		0.00	2	0.00	1	0.00	0	0.00	0	0.00	0	0.0040	3	758
3' downstream region	99245373	32146 G>T	rs76871422	0.01	6	0.00	3	0.00	1	0.00	1	0.00	3	0.0187	14	750
3' downstream region	99245364	32155 C>T	rs57922842	0.03	21	0.02	17	0.03	21	0.01	7	0.02	15	0.1080	81	750
3' downstream region	99245311	32208 G>T		0.01	6	0.00	3	0.01	4	0.01	8	0.00	3	0.0320	24	750
3' downstream region	99245280	32239 T>G	rs4646457	0.07	53	0.07	54	0.14	108	0.11	79	0.07	50	0.4587	344	750
3' downstream region	99245275	32244 T>C	rs4646456	0.20	147	0.18	135	0.20	148	0.19	140	0.19	143	0.9507	713	750
3' downstream region	99245267	32252 C>T		0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	750
3' downstream region	99245241	32278 A>G		0.00	0	0.00	3	0.00	0	0.00	0	0.00	0	0.0040	3	750
3' downstream region	99245160	32359 G>A		0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1	758
3' downstream region	99245013	32506 A>C	rs4646458	0.06	42	0.05	34	0.10	76	0.07	53	0.05	38	0.3240	243	750
3' downstream region	99244460	33059 GT microsatellite repeat	rs10536492	-	-	-	-	-	-	-	-	-	-	-	-	740
3' downstream region	99244392	33127 T>C	rs10259288	0.01	4	0.01	4	0.02	13	0.01	8	0.01	6	0.0464	35	754
3' downstream region	99244032	33487 T>C		0.00	1	0.00	0	0.00	0	0.00	0	0.00	0	0.0013	1	758

Figure 5.6: The number of times a particular variant is observed within the Ethiopian cohort (n=758 chromosomes). The frequency refers to the number of chromosomes on which a particular variant was identified. A "singleton" is a variant that was observed in a single heterozygous individual. The *y*-axis shows the proportion of the total amount of identified variation that is attributed to variants of particular frequencies; i.e. singletons account for ~30.3% of all identified polymorphic sites.



5.2.2 Analysis of variation observed in the proximal promoter of CYP3A5

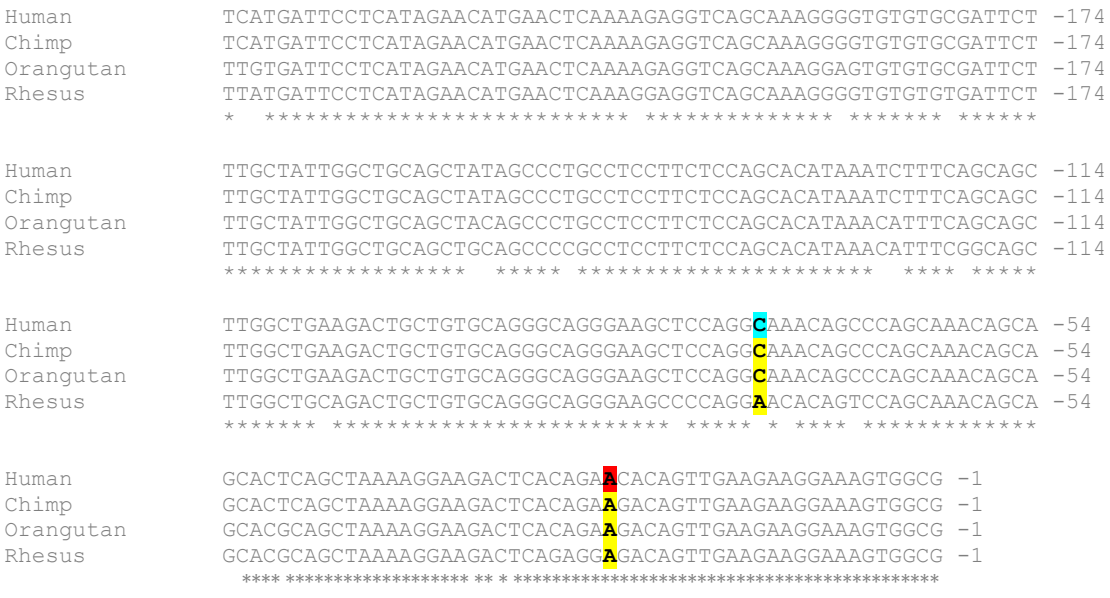
None of the variants observed in the *CYP3A5* promoter occur in experimentally established transcription factor binding sites. There is ~95% sequence identity between the human *CYP3A5* promoter sequence and the homologous primate sequences (see Table 5.7). A cross-species alignment of primate *CYP3A5* promoter sequences, Figure 5.7, found that 10/11 polymorphic sites were in highly conserved nucleotide positions in primates. A comparison of human *CYP3A5* sequence with the mouse (*Mus musculus)* found 51% homology, 60% with the cow (*Bos taurus*) and 61% in the pipid frog (*Xenopus tropicalis)*; indicating that *CYP3A5* sequence identity is highest between closely related primate species than to other distantly related species.

Table 5.7: A summary of the sequence similarity between the human *CYP3A5* promoter and the corresponding chimpanzee, Orang-utan and rhesus macaque sequences.

% sequence similarity	Human	Chimpanzee	Orang-utan	Rhesus macaque
Human	-			
Chimpanzee	99	-		
Orang-utan	93	93	-	
Rhesus macaque	92	92	89	-

Figure 5.7: An alignment of multiple primate *CYP3A5* promoter sequences performed using ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Stars indicate nucleotide positions which are identical in all four species. Positions are numbered from the ATG start codon where base A is +1. All identified human *CYP3A5* polymorphic sites which occur in highly conserved primate nucleotide positions are highlighted in red. Human variants which occur in non-conserved sites are highlighted in blue.

```
Human          ----------------------TCTATTGCTATCACCACAGAGTCAGAGGGGATGAGAC- -764
Chimp          -------------------------ATTGCTATCACCACAGAGTCAGAGGGGATGAGAC- -767
Orangutan      TTGAGTCCCAAGCAACCATTAGTCTATTGCTATCACCACAGAGTCAGAGGGGATGAGAC- -742
Rhesus         -----------------------CTATTGCTATCACCACAGAGTCAGAGGGGATGACACA -764
                                       ****************************** **

Human          -----GCCCAGCAATCTCACCCAAGACAACTCCACCAACATTCCTGGTTACCCACCATGT -709
Chimp          -----GCCCAGCAATCTCACCCAAGACAACTCCACCAACATTCCTGATTACCCACCATGT -712
Orangutan      -----GCCCAGGAATCTCACCCAAGACAACTCCACCAACACTCCTGGTTACCCACCGTGT -687
Rhesus         CGGGGGCCCAGCAATCTCACCCAAGTCAACTCCACCAACATTCCTGGTTACCCACCGTAT -704
                    ****** ************* ************** ***** ********* * *

Human          GTACAGTACCCTGCTAGGAACCAGGGTCATGAAAGTAAATAATACCAGACTGTGCCCTTG -649
Chimp          GTACAGTACCCTGCTAGGAACCAGGGTCATGAAAGTAA-TAATACCAGACTGTGCCCTTG -653
Orangutan      GTACAGTACCCTGCTAGGAACCAGGGTCATGAAAGTAAATAATACCAGACTGTGCCCTTG -627
Rhesus         GTACAGTACCCTGCTAGGGACCAGGGTCATGACAGTAAATAATACCAGACTGTGCCCTTG -644
               ***************** ************* ***** ********************

Human          AGGAGCTCACCTCTGCTAAGGGAAACAGGCATAGAAACTTACAATGGTGGTAGAGAGAAA -589
Chimp          AGGAGCTCACCTCTGCTAAGGGAAACAGGCACAGAAACTTACAATGGTGGTAGAGAGAAA -593
Orangutan      AGGAGCTCACCTCTACTAAGGGAAACAGGCACAGAAACCCACAATGGTGGTAGAGAGAAA -567
Rhesus         AGGAGCTCACCTCTGCTAAGGGAAACATGCACAGAAACCCACAATGGTGGCAGAGAGGAA -584
               ************** ************ *** ******  ********** ****** **

Human          AGAGGACAATAGGACTGTGTGAGGGGGATAGGAGGCACCCAGAGGAGGAAATGGTTACAT -529
Chimp          AGAGGACAATAGGACTGTGTGAGGGGGATAGGAGGCACCCAGAGGAGGAAATGGTTACAT -533
Orangutan      AGAGGACAATAGAACTGTGTGAGGGGGATAGGAGGCACCCAGAGGAGGAAATGGTTACAT -507
Rhesus         AGAGGACAATGGGACTGTGTGAGGGGGATAGGAGGCACCCAGAGGAGGAAATGG------ -530
               ********** * ****************************************

Human          TTGTGTGAGGAGGTTGGTAAGGAAAAATTTTAGCAGAAGGGGTCTGTCTGGCTGGGCTTG -469
Chimp          TTGTGTGAGGAGGTTGGTAAGGAAAAATTTTAGCAGAAGGGGTCTGTCTGGCTGGGCTTG -473
Orangutan      TTGTGTGAGGAGGTTGGTAAGGAAAAATTTTAGCAGAAGGGGTCTGTCTGGCTGGGGTTG -447
Rhesus         ----GTGAGGAGGTTGGTAAGGAAAGATTTTAACAGAAGGGGTCTGTCTGGCTGGACATG -474
                   ********************* ****** ********************   **

Human          GAAGGATACGTAGGAGTCATCTAGAGGGCACAGGTACACTCCAGGCAGAGGGAATTTCGT -409
Chimp          GAAGGATACGTAGGAGTCATCTAGAGGGCACAGGTACACTCCAGGCAGAGGGAATTGCGT -413
Orangutan      GAAGGATGTGTAGGAGTCATCTAGGGGGCACAGGTACACTCCACGGAGAGGGAATTGCGT -387
Rhesus         GAAGGACATGTAGGAGTCATCTAGAGGGCACAGGTACACTCCAGGCAGAGGGAATTGCAT -414
               ******   *************** ****************** * ********** * *

Human          GGGTAAAGATGTGTAGGTGTGGCTTGTGAGGATGGATTTCAATTATTCTAGAATGAAGGC -349
Chimp          GGGTAAAGATGTGTAGGTGTGGCTTGTGAGGATGGATTTCAATTATTCTAGAATGAAGGC -353
Orangutan      GGGTAAAGATGTGTAGGTGTGGCTTGTGAGGATGGATTTCAATTATTCTAGAATGAAGGC -327
Rhesus         GGGTAAAGATCTGTAGGTATGGCTTGTGAGGATGGATTTCAATTATTCTGGAATGAGGGC -354
               ********** ******* ****************************** ****** ***

Human          AGCCATGGAG----GGGCAGGTGAGAGGAGGGTTAATAGATTTCATGCCAATGGCTCCAC -293
Chimp          AGCCATGGAGACAGGGGCAGGTGAGAGGAGGGTTAATAGATTTCATGCCAATGGCTCCAC -293
Orangutan      AGCTATGGAGACAGGGGCAGGTGAGAGGAGGGTTAATAGATTTCATGCCAATGGCTCCGC -267
Rhesus         GGCCATGGAGACAAGGGCAGGTGAAAGGAGAGTTAACAGATTTCATGCCAATGGCTCTGC -294
                ** ******    ********** ***** ***** ********************  *

Human          TT-GAGTTTCTGATAAGAACCCAGAACCCTTGGACTCCCCGATAACACTGATTAAGCTTT -234
Chimp          TT-GAGTTTCTGATAAGAACCCAGAACCCTTGGACTCCCCGATAACACTGATTAAGCTTT -234
Orangutan      TT-GAGTTTCTGATAAGAACCC-------------------------TGATTAAGTTGT -234
Rhesus         TTTGAGTTTCTGATAAGAACCCAGAACCCTTGGACTCCCCACTAACACTGATTAAGGTTT -234
               ** ******************                            ******** * *
```



Given the importance of CYP3A5 in drug metabolism two previous studies have only characterised hepatic and intestinal transcription factor binding sites (Lin et al. 2002; Burk et al. 2004). CYP3A5 is the major extra-hepatic drug metabolising enzyme and predictive analysis was used to identify which of the *CYP3A5* polymorphisms were candidates for affecting transcription factor binding in tissues in which the enzyme is known to be expressed (Lamba et al. 2002). Bioinformatics analysis predicted that 4 of the 9 identified *CYP3A5* promoter variants were candidates for affecting transcription factor binding. There are problems associated with predicting functionally relevant transcription factor binding sites given the intricacy of binding complexes in different tissues (Jones and Swallow 2011). For this analysis, variants were only considered to be candidates for affecting transcription factor binding to the *CYP3A5* promoter if the identified site was predicted to bind transcription factors present in tissues in which CYP3A5 protein is expressed and also if the three MatInspector scores were ≥85%. A summary of MatInspector predictions is provided in Table 5.8.

Table 5.8: A summary of the MatInspector analysis of identified *CYP3A5* promoter variants. A detailed outline of the criteria used to choose variants as candidates for affecting transcription factor binding is presented in section 2.3.9.2.

Position on chromosome 7: (NCBI Build 132, February 2009)	*CYP3A5* variant and its position relative to the translation initiation codon (A of ATG is +1)	NCBI dbSNP database refSNP ID(s)	Reference sequence	Predicted to affect transcription factor to *CYP3A5* and in which tissue(s)?	MatInspector scores	Variant sequence	Predicted to alter transcription factor binding?	MatInspector scores
99278224	-705 3 base pair deletion	-	TGTGTACAGTACCCT	Yes: testis	Opt score: 98% Core similarity: 100% Matrix score: 99%	TGTGTA---TACCCT	Yes; site is predicted to be disrupted	
99278223	-704 A>G	-	ACAGTACCCTGC	Yes: testis	Opt score: 98% Core similarity: 100% Matrix score: 99%	ACAGTGCCCTGC	Yes site is predicted to be disrupted	Opt score: 86% Core similarity: 100% Matrix score: 86%
99278146	-627 G>A	-	CTAAGGGAAACAG	Yes: liver, prostate and testis	Opt score: 98% Core similarity: 100% Matrix score: 98%	CTAAGGAAAACAG	Yes site is reported to be disrupted	Opt score: 98% Core similarity: 100% Matrix score: 99%
99278070	-551 C>A	rs28365069	GAGGCACCCAGAG	No	-	GAGGCAACCAGAG	Predicted to create a binding site for kidney specific transcription factors	Opt score: 93% Core similarity: 100% Matrix score: 97%

5.2.3 Analysis of variation observed in the CYP3A5 coding region

A total of 39 variants were identified in the *CYP3A5* coding region (Figure 5.5). There is ~93% sequence homology between the human *CYP3A5* coding region re-sequenced in this study and the corresponding regions in other primates; see Table 5.9. An alignment of primate *CYP3A5* coding regions found that 28/39 polymorphic sites were in highly conserved primate nucleotide positions. The highest frequency variants observed were the *CYP3A5*3* and *CYP3A5*6* defining alleles. *CYP3A5*7* was observed in two heterozygous individuals. No polymorphisms were identified in consensus splice sites and bioinformatics analysis of all intronic polymorphisms, with BDGP (http://www.fruitfly.org/seq_tools/splice.html), did not identify any variants which were predicted to affect pre-mRNA splicing.

Table 5.9: A summary of the sequence similarity between the human, chimpanzee, Orang-utan and rhesus macaque *CYP3A5* coding regions

% Sequence similarity	Human	Chimpanzee	Orangutan	Rhesus macaque
Human	-			
Chimpanzee	99	-		
Orangutan	88	89	-	
Rhesus macaque	92	92	84	-

Of the 76 polymorphic sites, 8 (10.5%) occur within exons and of these 5 (6.58%) of all identified exonic polymorphisms are predicted to cause changes to the amino acid sequence. Not all non-synonymous substitutions are damaging to the structure or function of a protein; there are degrees of similarity between different amino acids and the replacement of an amino acid with a similar one is called a conservative replacement and unlikely to have a major effect on the protein function or structure. Non-conservative changes (or radical changes) are much more likely to affect the structure or function of the protein as they may alter the polarity, molecular weight and chemical composition (Graur et al. 2000).

The effect of each non-synonymous substitution on CYP3A5 protein was analysed using PolyPhen2 software (http://genetics.bwh.harvard.edu/pph2/) (see section 2.3.9.3). The results are presented in Table 5.10. Only three of the five identified non-synonymous mutations were predicted to be damaging to the CYP3A5 protein.

Table 5.10: The results of PolyPhen2 analysis of non-synonymous substitutions on CYP3A5 protein

Region of *CYP3A5*	Position on chromosome 7	*CYP3A5* variant and its position relative to ATG start codon	Amino acid change	PolyPhen2 predicted effect	Confidence of PolyPhen2 prediction
Exon 4	99270249	7270G>A	G77S	Probably damaging	Score: 0.993 Sensitivity:0.69 Specificity: 0.97
Exon 8	99261651	15868A>G	K266R	Benign	Score: 0.009 Sensitivity:0.97 Specificity: 0.76
Exon 9	99260502	17017C>T	R268Stop	Damaging	-
Exon 10	99258124	19395A>C	K342T	Possibly damaging	Score: 0.629 Sensitivity:0.87 Specificity: 0.92
Exon 11	99250381	27138A>G	V350M	Benign	Score: 0.026 Sensitivity:0.96 Specificity: 0.80

The significance of differences in the proportion of synonymous and non-synonymous variation can be measured by a codon-based Z test which compares the relative abundance of either type of polymorphism in a gene sequence and calculates the significance of the differences using a Bootstrap method described by (Nei and Kumar 2000). The results were non-significant for these analyses; (Z=0.961 and p=0.169) indicating that the proportion of non-synonymous variation observed in the gene does not significantly differ from the amount of identified synonymous variation.

However a paper in 2004 examining the total number of amino acid changes within and between primate lineages found that within 103 protein-coding genes (26,999 codons), 147 amino acid changes were observed (0.56%) (Kitano et al. 2004). However the results for *CYP3A5* within Ethiopia (5/502 codons; ~1%) are double that for a fraction of the total number of codons analysed in the original study. This suggests that there are more non-synonymous substitutions observed within the *CYP3A5* gene than for multiple protein-coding genes. This suggests that variants for low/non-expression of the protein, in addition to *CYP3A5*3, CYP3A5*6* and *CYP3A5*7*, may be selected for within Ethiopia (positive selection) and selection for removing them (purifying selection) is weak.

5.2.4 Analysis of variation observed in the 3´ untranslated region of CYP3A5

A single, high frequency, variant was identified in the 3´ UTR (rs15524); see Figure 5.5 and Table 5.6. This variant has been previously reported to be in high linkage disequilibrium with the *CYP3A5*3* defining G allele; although a study found that this variant alone has no effect on CYP3A5 protein activity (Busi and Cresteil 2005).

5.2.5 Analysis of the gene flanking sequence

A total of 9 and 16 variants were identified in the 5´ and 3´ regions re-sequenced either side of the *CYP3A5* gene, see Figure 5.5 and Table 5.6. This included a –GT microsatellite repeat. ~32.9% of all identified polymorphic sites occurred in genomic sequence flanking the *CYP3A5* gene. The proportion of variation identified in both the 5´ (9 variants/1691 fixed sites) and 3´ (16 variants/2370 fixed sites) gene flanking regions was similar to that observed in the gene sequence itself (51 variants/8100 fixed sites). No significant differences in the proportion of variation observed between the gene sequence and the 5´ gene flanking region (chi-square=0.0225, p=0.8807, 1 d.f.) or the 3´ gene flanking region (chi-square=0.07301, p=0.7870, 1 d.f.). A comparison of the *CYP3A5* sequence with larger flanking regions on either side of the gene may ascertain whether the proportion of variation within flanking regions is greater than that for an equivalent region of the gene.

5.3 Variation at the *CYP3A5* locus in non Ethiopian sub-Saharan Africans

As reported in chapter 4, a novel 10 base pair deletion, identified in intron 1, was not predicted to affect pre-mRNA splicing by BDGP (http://www.fruitfly.org/seq_tools/splice.html). However given the close proximity of this deletion to the intron 1 splice site, experimental evidence will provide an indication of how this deletion may affect *CYP3A5* transcription. No other identified polymorphism in the non-Ethiopian African cohort (see Table 4.2) was predicted to affect the initiation of *CYP3A5* transcription, mRNA processing or splicing.

5.4 Comparing intra-African *CYP3A5* diversity in a global context

To date, the full *CYP3A5* coding regions have been re-sequenced in five diverse populations (see Table 5.12), excluding Ethiopian groups within this study [(Thompson et al. 2004)(http://egp.gs.washington.edu/)]. The proximal promoter has been re-sequenced in African-Americans, Europeans and Han Chinese individuals (Thompson et al. 2004). Some re-sequencing data have also been generated as part of the 1000 Genomes database, however the coverage for the majority of individuals (~700) re-sequenced within this project is still relatively low at X2 (Zhang and Dolan 2010). Consequently rare variants are much more difficult to identify; since *CYP3A5* is characterised by an excess of rare variants (Thompson et al. 2004), the 1000 Genomes data are not ideal to compare with those generated for this study at present.

A total of 36 polymorphic sites were identified within a 13,853 base pair region of chromosome 7, including the *CYP3A5* promoter and coding region, re-sequenced in 24 African-Americans, 24 Europeans and 24 Han Chinese individuals (Thompson et al. 2004). 20/36 polymorphic sites were specific to African-American individuals. The European and Han Chinese individuals re-sequenced had high frequencies of the *CYP3A5*3* defining allele and a variant present in the 3′ UTR (rs15524) but had very few additional variants.

A total of 81 polymorphic sites were identified within a 17,222 base pair region of *CYP3A5* re-sequenced in 95 individuals from 5 different population groups. Populations with recent African ancestry had more variation at the *CYP3A5* locus than European, Hispanic and Asian populations.

The African data reported in this, and the previous, chapter are consistent with previous observations that there is more variation at the *CYP3A5* locus in populations with recent African ancestry than in other global populations. Sub-Saharan African groups, including those from Ethiopia, had more variation within a 4448 base pair region of *CYP3A5* than a 17,222 base pair region re-sequenced in East Asian and Hispanic populations from the NIEHS cohort. Africans also had more variation within a 4448 base pair region than was reported for a 13,853 base pair region re-sequenced in European and Han Chinese populations (Thompson et al. 2004). Across the overlapping 4448 base pair region, a total of 17 identified variants were unique to African populations; although it is interesting to note that the majority were singletons and none were predicted to have an effect on pre-mRNA splicing or gene transcription.

Table 5.12: A full list of all population groups in which *CYP3A5* has been re-sequenced. Information is provided on the number of variant sites identified in each group and what regions of the gene have been re-sequenced.

Population group	Number of individuals re-sequenced	Total amount of sequence (base pairs)	Regions of *CYP3A5* re-sequenced				Number of polymorphic sites identified	References
			Promoter	Exons	Introns	3´ UTR		
African-Americans	24	13,853	✓	✓	Flanking	✓	25	(Thompson, Kuttab-Boulos et al. 2004)
Europeans	24	13,853	✓	✓	Flanking	✓	13	(Thompson, Kuttab-Boulos et al. 2004)
Han Chinese	24	13,853	✓	✓	Flanking	✓	7	(Thompson, Kuttab-Boulos et al. 2004)
African-Americans	15	17,222	Partly	✓	All***	✓	37	(http://egp.gs.washington.edu/data/cyp3a5/)
Yoruba in Ibadan, Nigeria	12	17,222	Partly	✓	All***	✓	40	(http://egp.gs.washington.edu/data/cyp3a5/)
Europeans	22	17,222	Partly	✓	All***	✓	23	(http://egp.gs.washington.edu/data/cyp3a5/)
Hispanic	22	17,222	Partly	✓	All***	✓	18	(http://egp.gs.washington.edu/data/cyp3a5/)
East Asians	24	17,222	Partly	✓	All***	✓	16	(http://egp.gs.washington.edu/data/cyp3a5/)

N.B. All*: some *CYP3A5* intronic regions have not been re-sequenced in NIEHS SNPs populations. The flanking introns have been re-sequenced in all populations although some additional intronic sequence information is missing.

5.4 Discussion

5.4.1 Variation in the CYP3A5 gene in diverse Ethiopian populations

Multiple variants were identified across the *CYP3A5* locus in thirteen African populations (also seen in chapter 4). An interesting observation is that ~30.3% of all variants identified within the 12,237 base pair region, re-sequenced in five Ethiopian populations, are singletons. Within the gene region itself ~40% of all identified Ethiopian variants are singletons. Additionally the proportion of non-synonymous variants was unusually high and twice the frequency for a much larger protein-coding genomic region (Kitano et al. 2004). Normally protein-coding regions and regulatory regions have functional constraint and low tolerance of mutations which can cause low/non-expression. The intensity of purifying selection (which removes damaging mutations) is determined by the degree of intolerance of a particular genomic site towards deleterious mutations (Graur et al. 2000). Studies of multiple genes have often found that genes which have high rates of protein evolution, or are losing their function (such as pseudogenes), have higher levels of non-synonymous mutations than conserved genes (Graur et al. 2000; Bachtrog 2008). The high frequencies of *CYP3A5*3* in the Afar, Amhara, Oromo and Maale, coupled with an excess of rare variants and high frequencies of non-synonymous variation are consistent with signatures of a selective sweep at the *CYP3A5* locus in Ethiopia. A comparison of the 4448 base pair region between all thirteen sub-Saharan African groups (see chapter 4) found that there were more rare variants within the Ethiopian cohort than in other populations from the sub-continent. This may be evidence of different regional selective pressures within sub-Saharan Africa, or a result of differences in population histories.

5.4.2 A number of identified novel variants are predicted to affect CYP3A5 transcription and protein expression in Africans

An intronic novel ten base pair deletion, immediately adjacent to the 3´ splice site of exon 1, was identified in five heterozygous individuals from West Central Africa (see Table 4.2) and may have implications for gene transcription and protein expression. An *in vitro* splicing assay such as the exon-trapping technique (Webb et al. 2003) can be performed to resolve the issue. Briefly, a double stranded DNA fragment, known to contain the ten base pair deletion, is cloned into a bacterial vector, transformed into bacterial culture and then isolated and transfected into eukaryotic cells. Following incubation and extraction of mRNA, and reverse transcription to cDNA, an examination and comparison of the cDNA sequence from

the clones containing the ten base pair deletion with a clone containing the ancestral sequence to see how and whether the sequences differ.

Four identified promoter variants were predicted to have an effect on gene transcription in Ethiopians. It is important to note that there are significant drawbacks to forming conclusions from bioinformatics analysis alone. Prediction programs generate a large number of results, of which the majority are false positives. Transcription factor binding sites vary considerably in length (~6-20 base pairs), are imprecise and can often bind many different transcription factors (Lapidot et al. 2008). Although transcription factor binding site motifs can vary, there are some bases which have a high probability of occurring within the motifs (Lapidot et al. 2008). Prediction programmes, such as MatInspector, compare the observed sequence for motifs and match them against a database of known regulatory motifs to see if these common bases occur within the sequence and so are characteristic of a binding site. The probability of the input sequence being similar to common binding site motifs is calculated. This is called a position weight matrix where the position of a particular nucleotide (of the input sequence) is compared with known regulatory motifs to calculate the likelihood that it is a true transcription factor binding site (Lapidot et al. 2008). However, because of the flexibility in binding site specificity, mutations which occur in regulatory motifs are not necessarily going to alter the affinity for a transcription factor to a particular site. Additionally, transcription factors may only influence gene transcription in particular cells and polymorphic sites within particular motifs may be cell-type specific (Jones and Swallow 2011). MatInspector provides tissue information for inferred binding sites, and so the multiple results can be filtered to identify candidate transcription factor binding sites.

However, only experimental methods can provide tangible evidence of the effects of identified polymorphisms on transcription factor binding. Electrophoretic mobility shift assays (EMSAs) are a straightforward way to determine the effect of nucleotide substitutions on transcription factor binding. Briefly, EMSA is a technique used to characterise protein/DNA interactions. DNA molecules are negatively charged and will migrate rapidly towards a positive electrode under an electric field, protein molecules which are bound to DNA will cause it to migrate slower through a polyacrylamide gel; the larger the protein bound to DNA the slower it moves through an electric field. For this assay, a short DNA fragment, of specific sequence and length, is radioactively labelled and incubated with purified proteins and non-specific DNA competitors. Following incubation the complex is analysed on a polyacrylamide gel, a positive interaction between the radiolabelled DNA and a protein will result in a shift in the mobility of the DNA fragment, as determined by a comparison with a radiolabelled free DNA molecule as a control (Alberts 2002).

EMSAs are often performed alongside DNase footprinting assays; which identify the DNA sequence a protein binds to. Briefly, a radiolabelled protein-bound DNA fragment, of known sequence and length, is incubated with a non-specific DNA cleaving restriction enzyme. The protein protects the DNA motif from cleavage through strong binding to the site. The complex is then run on a polyacrylamide gel and a gap, or footprint, is seen where the restriction enzyme has not been able to cut; thus identifying the DNA motif that the protein binds to (Alberts 2002).

Both EMSA and DNA footprinting techniques will provide evidence that a DNA fragment is capable of binding a specific protein *in vitro* and will identify the sequence motif. However it is important to note that *in vivo* binding is often complicated and a particular protein may not preferentially bind to the identified motif *in vivo*.

An interesting observation is that three of the four promoter variants, which are candidates for affecting *CYP3A5* transcription, occur at frequencies of ≤1%; the novel ten base pair deletion, likely to have an effect on pre-mRNA splicing in five West Central African individuals, was observed at a frequency of ~0.007%; and all identified non-synonymous mutations were singletons. Although the proportion of all variants identified which were exonic variants was high (~10.5%) and of these 6.58% were non-synonymous changes. Of the nine previously reported candidates for causing polymorphic CYP3A5 expression (Table 5.4); only *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* were identified in the African re-sequencing data. This suggests that *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* are the main determinants of polymorphic CYP3A5 expression in all global populations that have been re-sequenced to date.

5.4.3 Comparing African CYP3A5 variability with other global populations

As expected there is more variability in the *CYP3A5* gene in populations with recent African ancestry than in other global populations. This is consistent with previous reports that genetic diversity within populations with recent African ancestry is greater than in non-African populations (Tishkoff et al. 2009). A comparison of *CYP3A5* variability between sub-Saharan African populations and African-Americans found that there was more variation in sub-Saharan Africa (discussed further in chapters 4 and 6). Considerable European ancestry within African-American populations may be influencing the amount of variation observed in the *CYP3A5* gene in this population (Reed 1969; Destro-Bisol et al. 1999). Populations with high frequencies of the derived *CYP3A5*3* allele had an excess of rare variants, supporting a hypothesis of directional selection within these populations. Interestingly, large numbers of

rare variants were observed in all Ethiopian groups, except the Anuak. This may be evidence of differential selection on *CYP3A5* within parts of Africa. Considerable admixture within North Eastern Ethiopian populations with those from the Arabian Peninsula (De Stefano et al. 2002; Lovell et al. 2005) may also be influencing the pattern of *CYP3A5* variability observed in the Afar, Amhara and Oromo.

The results from this chapter indicate that population structure at the *CYP3A5* gene in diverse Ethiopian populations is likely to be influenced by novel, rare variants in addition to *CYP3A5*3, CYP3A5*6* and *CYP3A5*7*. The following chapter will examine intra-Ethiopian differences alongside re-sequencing data for Coriell populations. The recent evolutionary history of *CYP3A5* will be examined in detail, including evidence of selection on the gene, in chapter 7.

6. Analysing intra-Ethiopian diversity at the *CYP3A5* gene in a global context

A number of polymorphic sites, identified in chapter 5, are novel candidates for causing low/non CYP3A5 expression phenotypes in Africa. Previously (Gebeyehu et al. 2011) reported collective frequencies of *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* in Ethiopians from multiple ethnic groups, not accounting for inter-ethnic diversity. It has been consistently shown in this thesis that the Ethiopian Afar, Amhara and Oromo significantly differ from other Africans and are characteristic of both African and non-African populations. This chapter aims to examine intra-Ethiopian diversity across the entire *CYP3A5* gene in detail and examine the observed diversity alongside three ethnically diverse Coriell populations from North America. Statistical associations between all identified polymorphic sites will be analysed through haplotype and linkage disequilibrium analyses. In addition, analyses of the allele frequency spectrum to test for significant departures from neutrality are presented.

6.1 *CYP3A5* variation in Ethiopia

6.1.1. Haplotype association of Ethiopian variants

Haplotypes were inferred from genotype data to examine the distribution of identified variants on Ethiopian chromosomes. Singleton variants, which could not be inferred onto a specific haplotype, were omitted from this part of the analysis. 100 haplotypes were inferred for a 12,237bp region, see Figures 6.1a-e, and ~81% of these occurred at a frequency of <1%. ~55% of all inferred haplotypes were defined by the *CYP3A5*3* mutation alone, ~17.7% by *CYP3A5*6* and ~0.003% by *CYP3A5*7.* The remaining haplotypes were all defined by the ancestral *CYP3A5*1* allele. *CYP3A5*3* defined the majority of haplotypes observed in the Afar, Amhara and Oromo, and *CYP3A5*1* defined the majority in the Maale, and nearly all in the Anuak (Figure 6.2).

The modal *CYP3A5*3* haplotype was ~7x more frequent than the second most frequent within this haplogroup. There were two main *CYP3A5*6* haplotypes; one ~2x more frequent than the other. The modal *CYP3A5*3* and *CYP3A5*6* haplotypes are characterised by high levels of polymorphisms in the *CYP3A5* flanking regions and a paucity of variation within the coding region or promoter. This is also true for the one *CYP3A5*7* haplotype.

All non-synonymous and promoter variants predicted to affect gene transcription and protein translation (reported in chapter 5), occur on *CYP3A5*1* haplotype backgrounds. Each identified non-synonymous polymorphism was a singleton; as a result not all exonic polymorphisms could be inferred onto specific *CYP3A5*1* haplotypes.

Figure 6.1a: Inferred *CYP3A5*3* haplotypes for five Ethiopian populations. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. The position defining *CYP3A5*3* is shown in red. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The numbers at the end of each row correspond to the number of Ethiopian chromosomes of a particular haplotype. Each haplotype was assigned a code, as shown in the column called "CODE".

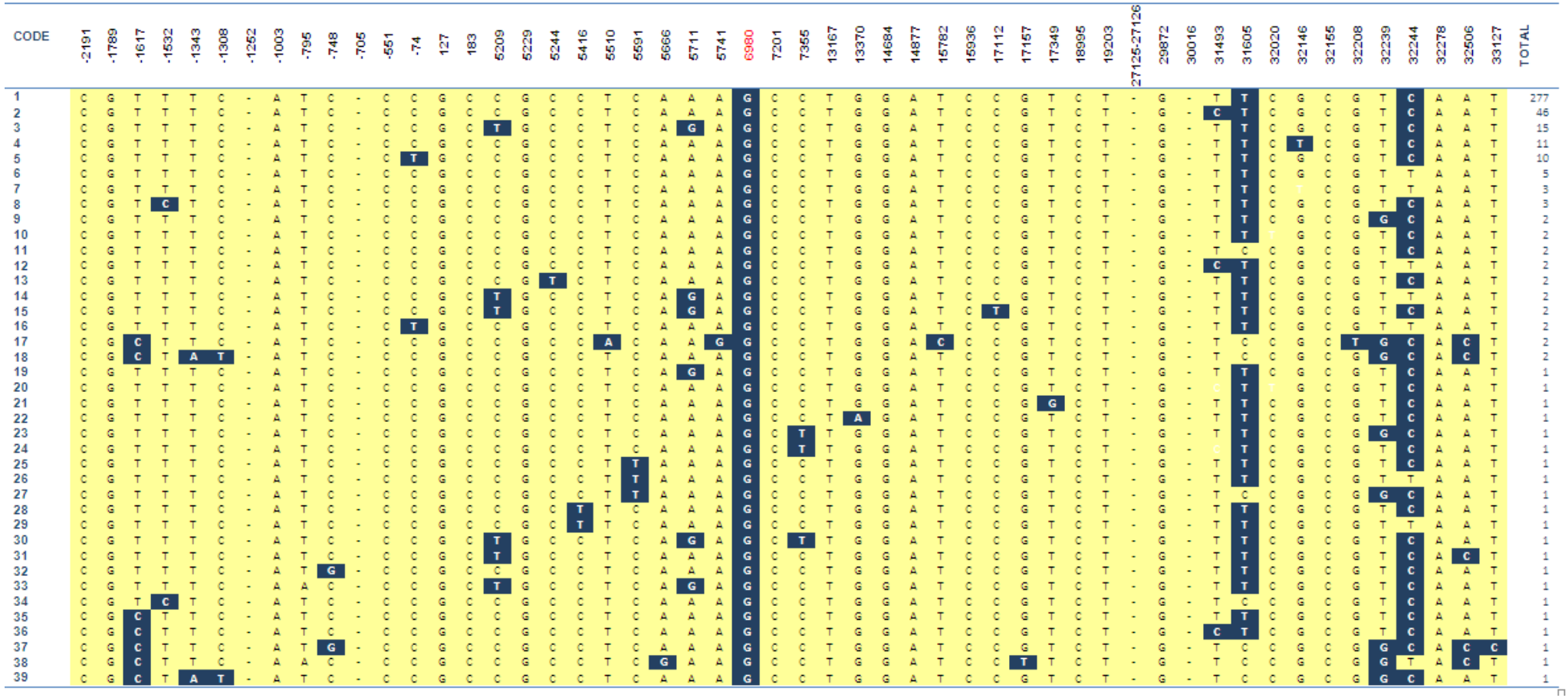
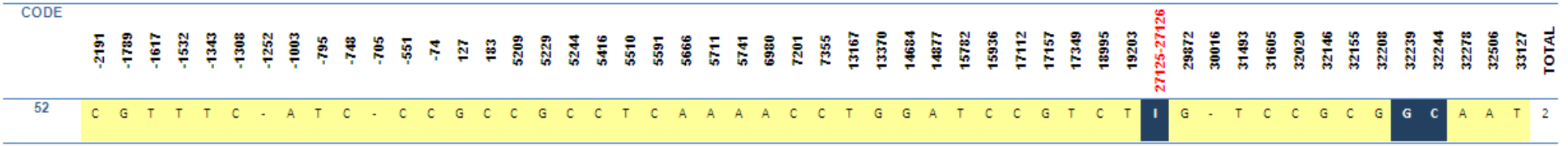


Figure 6.1b: Inferred *CYP3A5*6* haplotypes for five Ethiopian populations. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. The position defining *CYP3A5*6* is shown in red. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The numbers at the end of each row correspond to the number of Ethiopian chromosomes of a particular haplotype. Each haplotype was assigned a code, as shown in the column called "CODE".

CODE	-2191	-1789	-1617	-1532	-1343	-1308	-1252	-1003	-795	-748	-705	-551	-74	127	183	5209	5229	5244	5416	5510	5591	5666	5711	5741	6980	7201	7355	13167	13370	14684	14877	15782	15936	17112	17157	17349	18995	19203	27125-27126	29872	30016	31493	31605	32020	32146	32155	32208	32239	32244	32278	32506	33127	TOTAL
40	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	T	G	G	C	A	C	T	74
41	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	G	C	A	C	T	40
42	T	G	T	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	T	G	G	C	A	C	T	2
43	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	T	G	G	C	A	A	T	2
44	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	C	-	G	-	T	C	C	G	C	G	G	C	A	C	T	2
45	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	T	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	G	C	A	C	T	2
46	C	G	T	T	T	C	-	C	T	C	-	A	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	D	T	T	C	G	C	G	G	C	A	A	T	1
47	C	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	G	C	A	C	T	1
48	T	G	T	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	G	C	A	C	T	1
49	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	T	T	A	A	T	1
50	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	T	T	A	C	T	1
51	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	T	G	T	T	A	C	T	1

Figure 6.1c: Inferred *CYP3A5*7* haplotypes for five Ethiopian populations. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. The position defining *CYP3A5*7* is shown in red. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The numbers at the end of each row correspond to the number of Ethiopian chromosomes of a particular haplotype. Each haplotype was assigned a code, as shown in the column called "CODE".

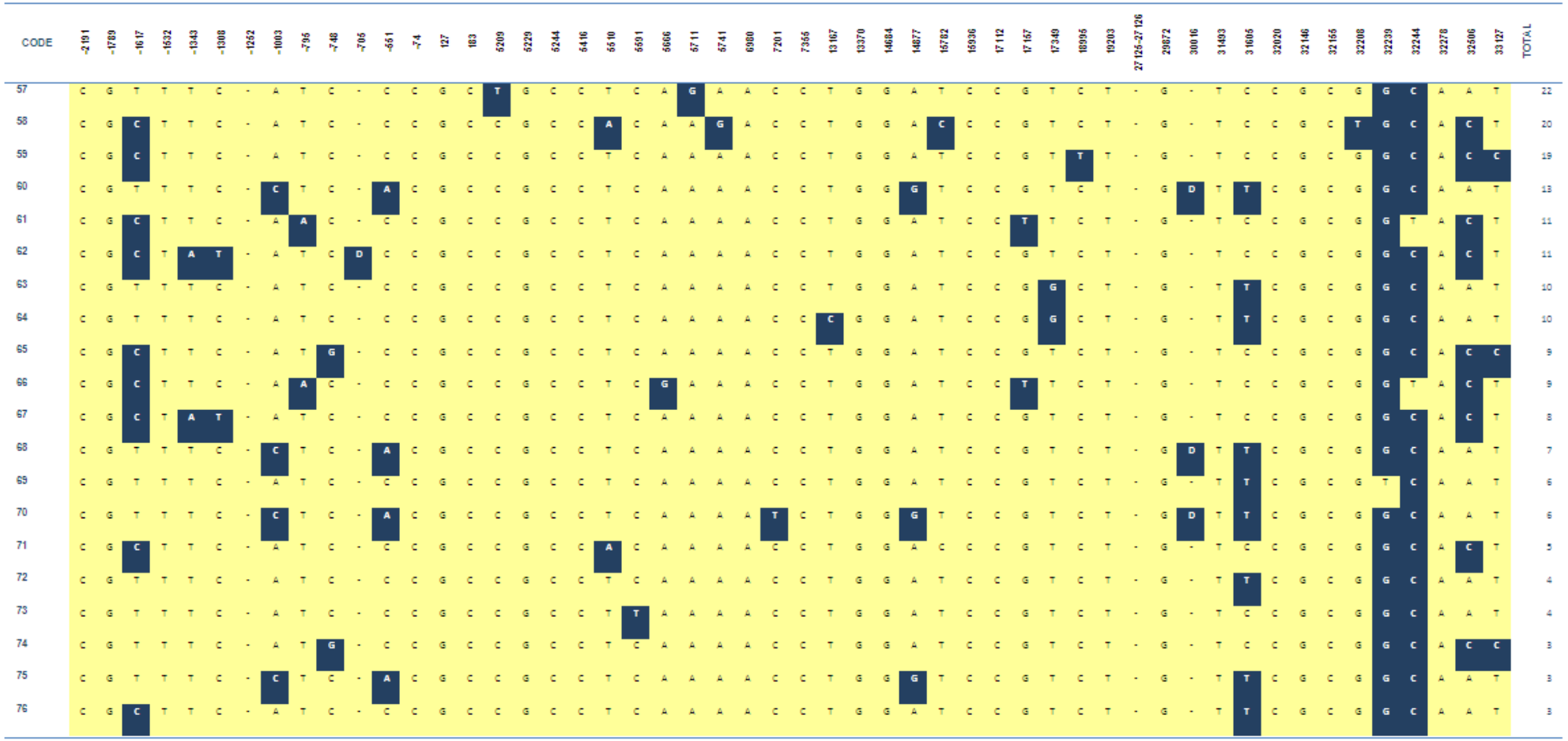


CODE	-2191	-1789	-1617	-1532	-1343	-1308	-1252	-1003	-795	-748	-705	-551	-74	127	183	5209	5229	5244	5416	5510	5591	5666	5711	5741	6980	7201	7355	13167	13370	14684	14877	15782	15936	17112	17157	17349	18995	19203	27125-27126	29872	30016	31493	31605	32020	32146	32155	32208	32239	32244	32278	32506	33127	TOTAL
52	C	G	T	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	A	C	C	T	G	G	A	T	C	C	G	T	C	T	I	G	-	T	C	C	G	C	G	G	C	A	A	T	2

Figure 6.1d: Inferred *CYP3A5*3/*6* recombinant haplotypes for five Ethiopian populations. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. The position defining *CYP3A5*3* and *CYP3A5*6* are shown in red. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The numbers at the end of each row correspond to the number of Ethiopian chromosomes of a particular haplotype. Each haplotype was assigned a code, as shown in the column called "CODE".

CODE	-2191	-1789	-1617	-1532	-1343	-1308	-1252	-1003	-795	-748	-705	-551	-74	127	183	5209	5229	5244	5416	5510	5591	5666	5711	5741	6980	7201	7355	13167	13370	14684	14877	15782	15936	17112	17157	17349	18995	19203	27125-27126	29872	30016	31493	31605	32020	32146	32155	32208	32239	32244	32278	32506	33127	TOTAL
53	C	G	T	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	G	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	T	C	G	C	G	T	C	A	A	T	3
54	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	G	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	C	G	G	C	A	C	T	1
55	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	G	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	T	G	G	C	A	A	T	1
56	T	G	C	T	T	C	-	A	T	C	-	C	C	G	C	C	G	C	C	T	C	A	A	A	G	C	C	T	G	A	A	T	C	C	G	T	C	T	-	G	-	T	C	C	G	T	G	G	C	A	C	T	1

Figure 6.1e: Inferred *CYP3A5*1* haplotypes for five Ethiopian populations. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The numbers at the end of each row correspond to the number of Ethiopian chromosomes of a particular haplotype. Each haplotype was assigned a code, as shown in the column called "CODE".



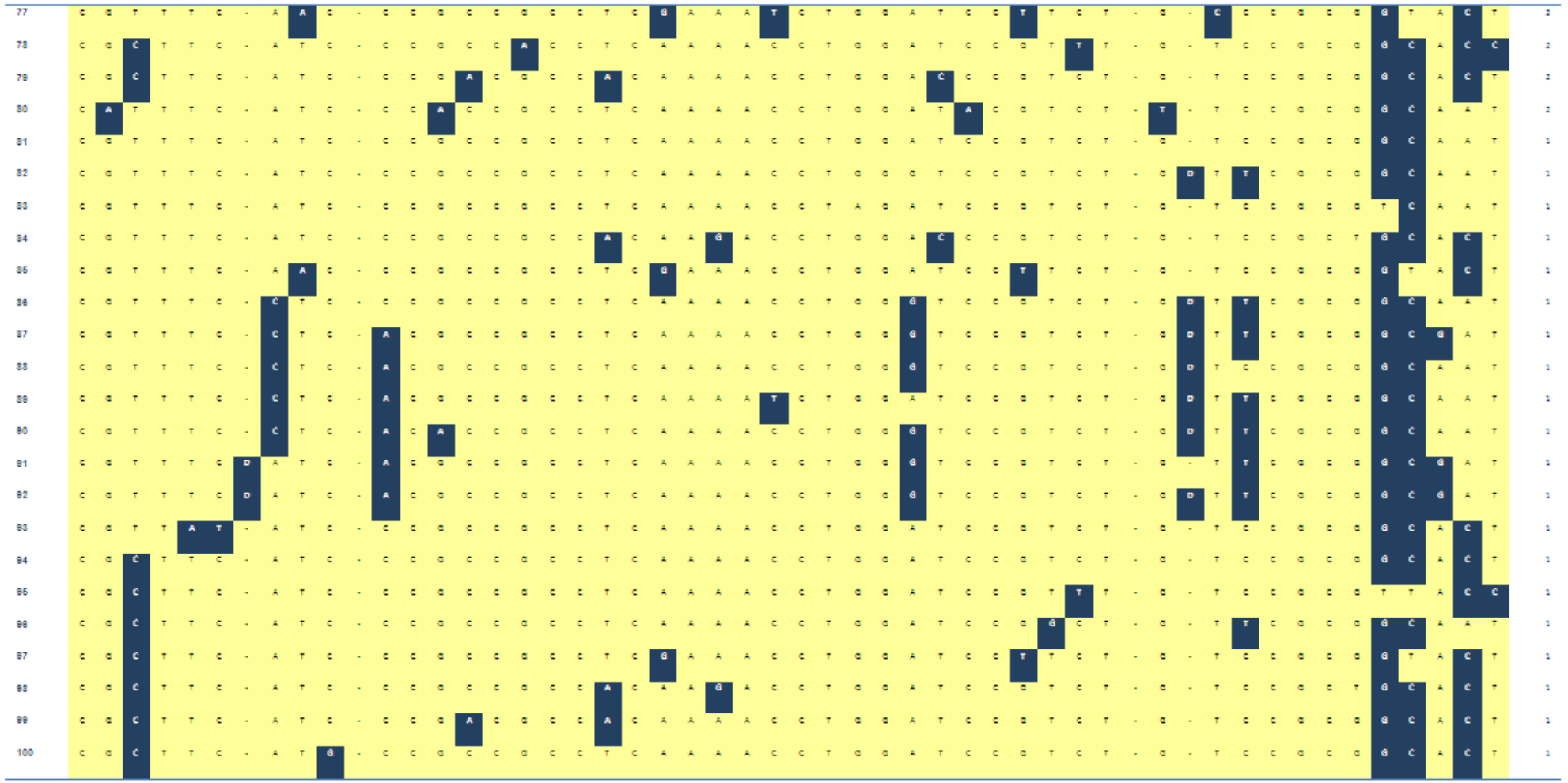


Figure 6.2a: The frequency of *CYP3A5* haplotypes (%) within each Ethiopian population. The key is numbered according to the haplotype codes assigned in Figures 6.1a-e; the numbers in brackets indicate which *CYP3A5* haplogroup (*CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*3/*6* or *CYP3A5*7*) each haplotype belongs to.

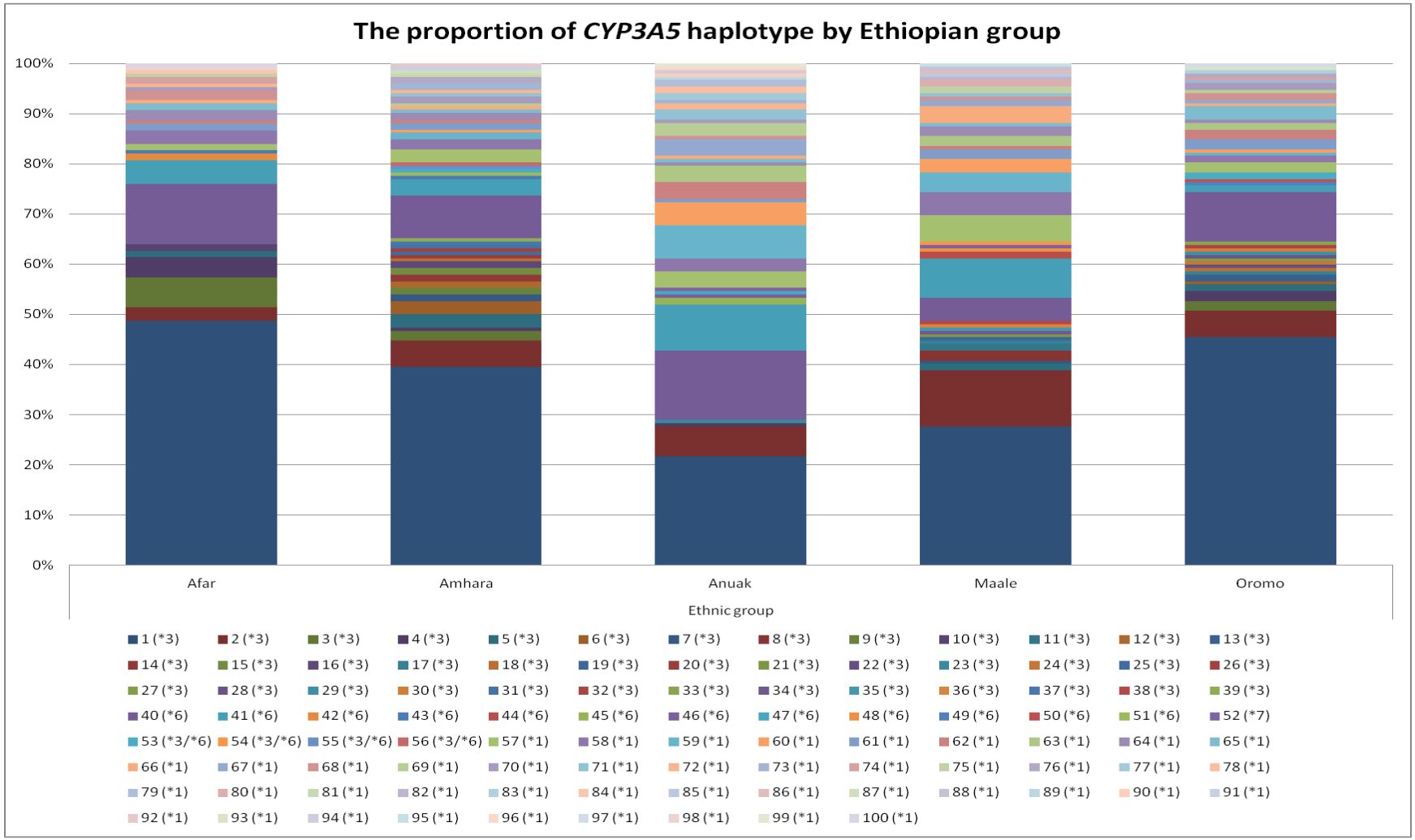
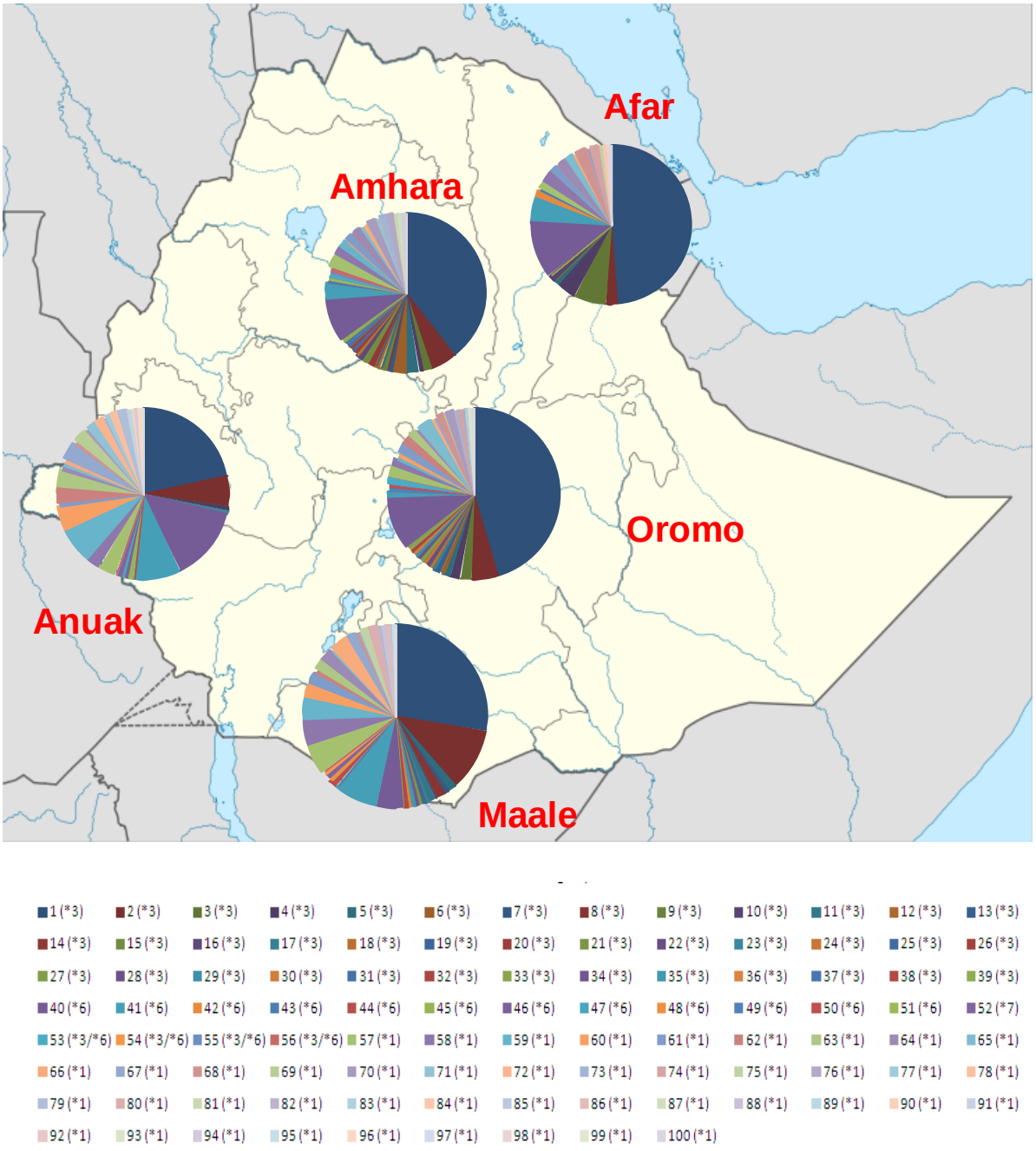


Figure 6.2b: A map showing the spatial distribution, of *CYP3A5* haplotypes, across Ethiopia. The map image has been taken from: http://commons.wikimedia.org/wiki/File:Ethiopia_location_map.svg?uselang=fr, and has been adapted by me in Microsoft PowerPoint 2007.



An interesting observation is that levels of diversity in the *CYP3A5*3* haplogroup are highest in populations closest to the Arabian Peninsula. *CYP3A5*7* haplotypes are observed at low frequencies in Ethiopia and are also found in the two populations located furthest from the Arabian Peninsula (Anuak and Maale).

6.1.2 Examining linkage disequilibrium across CYP3A5

Pairwise linkage disequilibrium (LD), calculated using the *D´* parameter, was measured within Ethiopia as a whole and within each of the five populations. Genotype data for all polymorphic loci, except singletons, were included for LD analyses. The results are consistent with those for haplotype inference. Within Ethiopia, the majority of loci are in complete LD (*D´*=1), see Supplementary Figure 3 on CD, consistent with previous studies (Thompson et al. 2004; Thompson et al. 2006) although not all associations are statistically significant. High levels of non-significant LD are a result of an excess of low frequency variants in the dataset. These variants occur on a single haplotype background but, due to their low frequencies, are not significantly associated. However *D´*=1 indicating that there is no recombination between specific loci, even if the association is not statistically significant.

There are distinct LD patterns, relating to the *CYP3A5*3* and *CYP3A5*6* variants, see Table 6.1. Across the 12,237bp region, *CYP3A5*3* and *CYP3A5*6* are in high LD with variants either side of the gene region. LD extends across a region of up to 35,318bp (bp); between variants located 2191bp upstream of the ATG start codon; and those located 2385bp downstream of exon 13.

D´ between *CYP3A5*3* and *CYP3A5*6* is equal to 1 and the association between the two loci is statistically significant. The *CYP3A5*6* allele almost always occurs on a haplotype with a *CYP3A5*1* allele. The high LD between *CYP3A5*3* and a variant at position 31605 (rs15524) and another at position 32244 (rs4646456) reflects the most common *CYP3A5*3* Ethiopian haplotype. The high LD between *CYP3A5*6* and variants in the 5´ [rs10270499 and rs776741] and 3´ [rs57922842, rs4646456, rs4646457, and rs4646458] flanking regions also reflect the most common *CYP3A5*6* haplotype observed in Ethiopia. *CYP3A5*7* allele frequencies are low and therefore very little can be deduced about LD from this locus alone. The results show that LD across *CYP3A5* is likely to extend beyond the gene region.

LD between *CYP3A5*7* and other polymorphic sites is insignificant (Table 6.1). This is almost certainly because the variant is observed at low frequency within Ethiopia.

Table 6.1: Pairwise LD between each of *CYP3A5*1/CYP3A5*3* (rs776746), *CYP3A5*6* (rs10264272) and *CYP3A5*7* (rs41303343) defining loci and all identified polymorphic sites across the Ethiopian cohort. Statistically significant *D´* values, following Bonferroni correction (adjusted *p*-value = 0.00096; for 52 tests), are highlighted in green.

Position on chromosome 7	*CYP3A5* region	Position of *CYP3A5* variant relative to ATG start codon	Pairwise *D´* values		
			rs776746 (99270539)	rs10264272 (99262835)	rs41303343 (99250397)
99279710	**5´ flanking region**	-2191	**0.944**	**1**	1
99279308		-1789	0.992	1	1
99279136		-1617	**0.896**	**0.923**	1
99279051		-1532	0.999	1	1
99278876		-1343	**0.649**	0.033	1
99278862		-1308	**0.649**	0.033	1
99278827		-1252	1	1	1
99278522		-1003	**1**	0.737	1
99278314	**Proximal promoter**	-795	**0.894**	0.542	1
99278267		-748	0.599	0.071	1
99278224		-705	**1**	0.158	1
99278070		-551	**1**	0.969	1
99277593		-74	1	0.736	1
99277392	**Intron 1**	127	1	1	1
99277337		182	1	1	1
99272310	**Intron 2**	5209	0.09	0.962	1
99272290		5229	1	1	1
99272275		5244	0.999	1	1
99272103	**Intron 3**	5416	0.999	0.224	1
99272009		5510	**0.914**	0.973	1
99271928		5591	0.22	0.467	1
99271853		5666	0.778	0.004	1
99271808		5711	0.122	0.962	1
99271778		5741	**0.881**	0.455	1
99270539		6980	-	**0.91**	0.992
99270504		7201	**1**	0.27	1
99270164	**Intron 4**	7355	1	1	1
99264352	**Intron 5**	13167	1	0.368	1
99264149	**Intron 6**	13370	0.09	1	1
99262835	**Exon 7**	14684	**0.91**	-	1
99262642	**Intron 7**	14877	**1**	0.837	1
99261737		15782	**0.909**	0.995	1
99261583	**Intron 8**	15936	0.992	1	1
99260407	**Intron 9**	17112	1	1	1
99260362		17157	**0.894**	0.542	1
99260170		17349	**1**	0.593	1
99258524		18995	**1**	0.998	1
99258316		19203	0.992	1	1
99250397	**Exon 11**	27125-27126	0.992	1	-
99247647	**Intron 12**	29872	0.992	1	1
99247503		30016	**1**	0.994	1
99246026		31493	**0.938**	**1**	1
99245914	**3´ untranslated region**	31605	**0.902**	**0.956**	0.999
99245499	**3´ flanking region**	32020	0.994	0.111	1
99245373		32146	**1**	1	1
99245364		32155	**0.937**	**1**	1
99245311		32208	**0.881**	0.455	1
99245280		32239	**0.94**	**0.895**	0.991
99245275		32244	0.143	0.264	1
99245241		32278	1	0.111	1
99245013		32506	**0.923**	**0.923**	1
99244392		33127	**0.92**	0.737	1

6.1.3 The CYP3A5 allele frequency spectrum and analyses for departures from neutrality

A total of 76 polymorphic sites were identified in 379 individuals from five Ethiopian populations (chapter 5). A summary of the number of polymorphic sites, and singleton variants, identified in each ethnic group is shown in Table 6.2.

Table 6.2: Molecular diversity estimates for the five Ethiopian populations. Percentages expressed in brackets indicate the proportion of variants, identified within a specific sample set, which are singletons.

Sample set	Number of identified polymorphic sites	Number of variants which occur on a single chromosome
Afar	37	8 (21.6%)
Amhara	41	6 (14.6%)
Anuak	49	17 (34.7%)
Maale	45	9 (20%)
Oromo	42	23 (57.8%)

The overall nucleotide diversity within the Ethiopian cohort is 4.7×10^{-4} and Tajima's *D* is -1.48; indicating a skew towards rare variants. Nucleotide diversity (π) varies between all five sample sets; see Table 6.3. The Afar, Amhara and Oromo all have lower measures of nucleotide diversity than the Anuak and Maale. Nucleotide diversity, in all five Ethiopian populations, was comparatively lower than for a similarly sized region in African-Americans (see Table 6.8).

Nucleotide diversity was examined for departures from neutrality using Tajima's *D*, Fu and Li's *D** and *F*,* and Fu's *FS* statistics, see Table 6.3. These methods evaluate whether there is an excess of rare variants within the re-sequenced region; a pattern consistent with a selective sweep or directional selection. All statistics reported negative values for each Ethiopian population, indicating a skew towards rare variants. However only Fu's *FS* reported a significant departure from neutrality, in all five Ethiopian populations, before and after Bonferonni correction ($p<0.0001$ for all populations).

The Afar, Amhara and Oromo have the largest skew towards rare variants. Interestingly, estimates for Tajima's *D* for these three populations are comparable to those for Han Chinese individuals (-1.21) (Thompson et al. 2004). Although the values suggest that the skew towards rare variants is not as high as reported for Europeans (-2.11) but much more than for African-Americans (-1.13).

Table 6.3: Molecular diversity indices for a 12,237bp *CYP3A5* region re-sequenced in five Ethiopian populations. All five Ethiopian groups showed a significant departure from neutrality, following Bonferonni correction (correction for 5 tests; $p \leq 0.01$), based on the Fu's *FS* *p*-value.

	Ethiopian sample set				
	Afar	**Amhara**	**Anuak**	**Maale**	**Oromo**
Nucleotide diversity (π)	3.91×10^{-4}	3.8×10^{-4}	5.14×10^{-4}	4.82×10^{-4}	3.55×10^{-4}
Tajima's *D*	-1.22	-1.57	-0.95	-0.75	-1.35
Fu and Li's *D**	-1.14	0.35	-2.09	0.65	0.25
Fu and Li's *F**	-1.41	-0.34	-1.93	0.08	-0.50
Fu's *FS*	**-12.13**	**-34.84**	**-19.68**	**-17.20**	**-24.74**

6.1.4 CYP3A5 haplotype diversity

The entire Ethiopian cohort (379 individuals; 758 chromosomes) was used to analyse haplotype diversity between inferred CYP3A5 expressers and low/non-expressers; and between the five *CYP3A5* haplotype classes: *CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*3/*6* and *CYP3A5*7*. The highest levels of gene diversity were observed in the Anuak (0.914 ± 0.012) and the lowest in the Afar (0.742 ± 0.036). Exact tests of population differentiation were performed to assess the significance of pairwise inter-population differences in gene diversity (Table 6.4). Consistent with the results from the geographic survey (chapter 3), the Afar, Amhara and Oromo do not significantly differ from each other, i.e. haplotype diversity estimates are similar in all three groups. The Anuak differ from all other Ethiopian sample sets; as do the Maale.

Within Ethiopia, gene diversity is highest in the Anuak, Maale and Amhara. The Anuak and Maale have the highest frequencies of *CYP3A5*1* haplotypes; which are the most diverse haplogroup (Table 6.5, Figure 6.1). This is likely to explain the high levels of gene diversity observed within these populations.

Figure 6.3: Haplotype diversity, calculated using Nei's h, for each of the five Ethiopian sample sets. Nei's *h* was calculated using Arlequin and the test_h_diff program written by Mike Weale (Thomas et al. 2002); results from both programmes were identical. Error bars denote standard deviation. Values above each data point show the Nei's *h* value for each sample set.

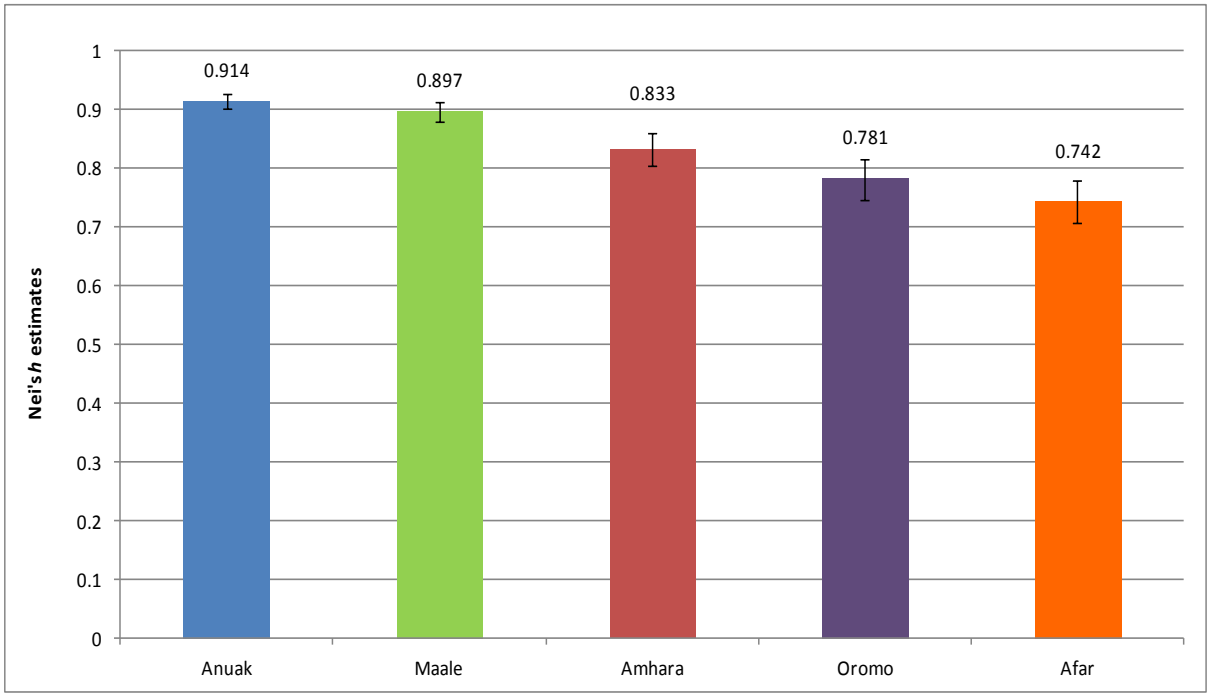


Table 6.4: Nei's h of the 12,237bp *CYP3A5* region across populations. Significance of pairwise differences, in diversity estimates, was measured using an exact test of population differentiation (executed in Arlequin software). *P*-Values, shown in the bottom left corner of the grid, which are significant after Bonferroni correction (for 5 tests; adjusted *p*-value = 0.01) are highlighted in green.

Nei's *h*	**0.742**	**0.833**	**0.914**	**0.897**	**0.781**
	Afar	**Amhara**	**Anuak**	**Maale**	**Oromo**
Afar	*				
Amhara	0.04	*			
Anuak	**<0.001**	**<0.001**	*		
Maale	**<0.001**	**<0.001**	**0.003**	*	
Oromo	0.40	0.72	**<0.001**	**<0.001**	*

Table 6.5: A comparison of heterozygosity in each of the five *CYP3A5* haplogroups: *CYP3A5*1*, *CYP3A5*3, CYP3A5*6, CYP3A5*3/CYP3A5*6* and *CYP3A5*7*. Significance was measured using the *h*-diff test (Thomas et al. 2002). Values which were significant after Bonferroni correction (for 5 tests; adjusted *p*-value = 0.01) are highlighted in green.

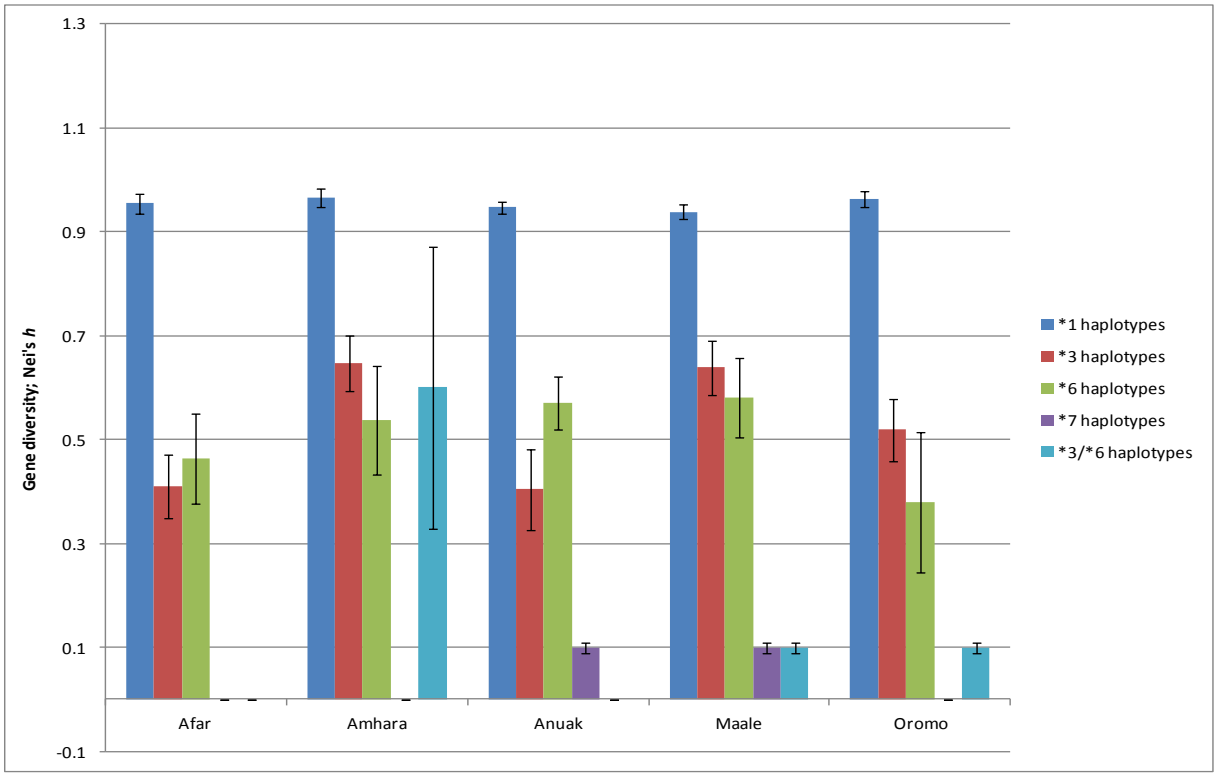
Nei's *h*	0.952	0.545	0.571	0.000	0.000
	CYP3A5*1	***CYP3A5*3***	***CYP3A5*6***	***CYP3A5*7***	***CYP3A5*3/*6***
CYP3A5*1	*				
CYP3A5*3	**<0.0001**	*			
CYP3A5*6	**<0.0001**	0.59	*		
CYP3A5*7	**<0.0001**	**<0.0001**	**<0.0001**	*	
CYP3A5*3/*6	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	*

The low/non-expresser haplogroups (*CYP3A5*3, CYP3A5*6, CYP3A5*3/*6* and *CYP3A5*7*) have significantly less diversity than *CYP3A5*1* haplotypes. One possible explanation is the *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* defining alleles are younger than the ancestral *CYP3A5*1.* New, and neutral, haplotypes generally fluctuate in their frequencies and have not had an equivalent amount of time to acquire additional mutations, compared to ancestral lineages.

The most common low/non-expresser *CYP3A5*3* and *CYP3A5*6* haplotypes, observed in Ethiopia, have very little variation over a large genomic region, i.e. they are characterised by extended haplotype homogeneity (Figures 6.1a-b). This may be a result of directional selection on particular *CYP3A5* variants (Sabeti et al. 2002). *CYP3A5*7* is at low frequency in Ethiopia and so not much can be inferred about selection on this allele; hence the low Nei's *h* calculated score. A recombinant *CYP3A5*3/CYP3A5*6* haplotype is observed in Ethiopia, although its low frequency suggests that *CYP3A5*3* and *CYP3A5*6* are subject to independent selective pressures and events.

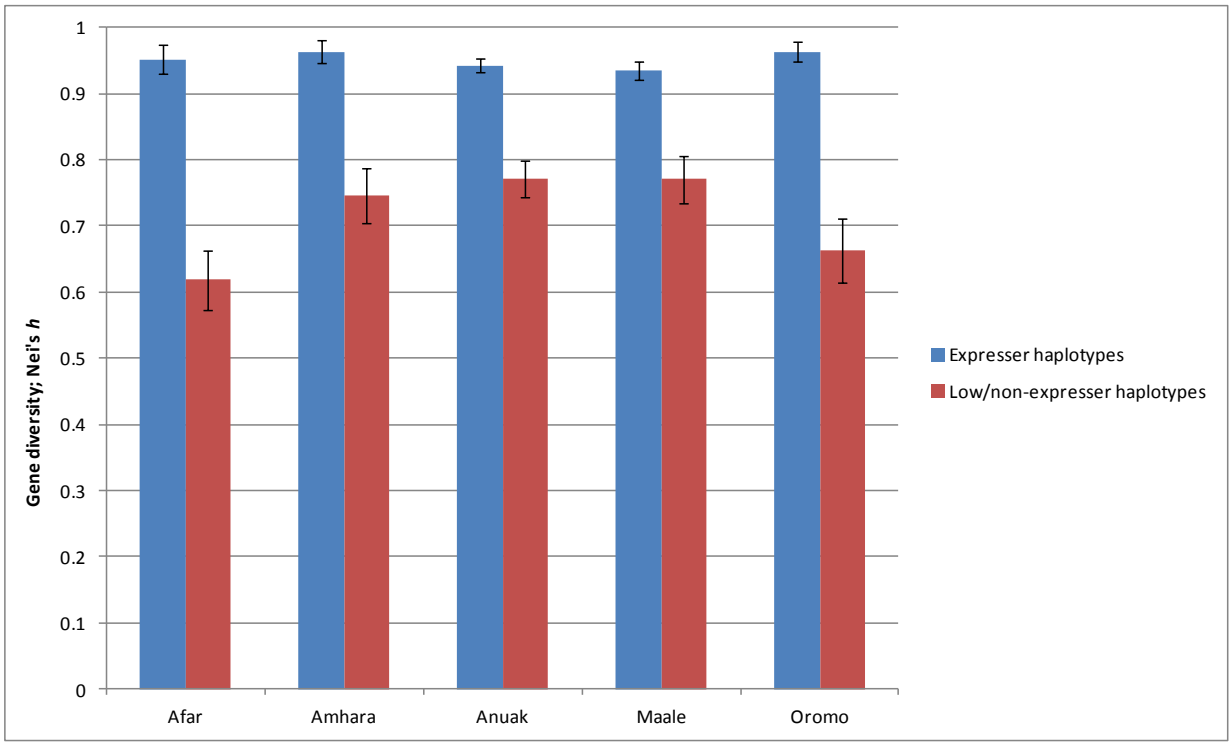
A comparison of diversity in each *CYP3A5* haplotype class, by Ethiopian group, is shown in Figure 6.4. *CYP3A5*7* and *CYP3A5*3/*6* recombinant haplotypes are observed at low frequencies within the dataset; and that there is a large variance around the estimates of Nei's *h*. *CYP3A5*1* haplotypes are the most diverse of all haplotype classes in all Ethiopian sample sets. *CYP3A5*3* and *CYP3A5*6* are considerably less diverse than *CYP3A5*1* haplotypes across the Ethiopian cohort.

Figure 6.4: A comparison of diversity in each of the five *CYP3A5* haplotype classes (*CYP3A5*1*, *CYP3A5*3, CYP3A5*6, CYP3A5*3/*6* and *CYP3A5*7*) by Ethiopian sample set. Nei's *h* was calculated using the *h*-diff sets of function and executed in the R-programming environment (Thomas et al. 2002). Error bars denote standard deviation.



Nei's *h* was used to compare diversity between *CYP3A5* expresser and low/non-expresser haplotypes within each of the five populations, see Figure 6.5. Of the novel mutations predicted to affect *CYP3A5* gene transcription and/or protein translation only one could be inferred onto a particular (*CYP3A5*1*) haplotype. However, as outlined in section 5.4.2, bioinformatics software can only predict whether a polymorphism may have an effect on gene transcription, mRNA processing or protein translation; no variant can be conclusively classed as either expresser or low/non-expresser based on bioinformatics analysis alone. Therefore individuals carrying this variant were excluded from these analyses. Only variants which have been shown to affect CYP3A5 expression *in vitro/in vivo* were classified as low/non-expresser variants. This meant that, of all identified *CYP3A5* variants, only *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* were considered low/non-expresser and the remaining chromosomes as expresser.

Figure 6.5: A comparison of heterozygosity between *CYP3A5* expresser and low/non-expresser haplotypes in each Ethiopian sample set. Nei's *h* and the significance of differences were measured using the *h*-diff test (Thomas et al. 2002). All comparisons were significant following Bonferonni correction (for 5 tests; adjusted *p*-value = 0.01). Error bars denote standard deviation of Nei's *h*.



Within each population, significantly higher levels of diversity were observed on expresser haplotypes than on low/non-expresser haplotype backgrounds in all five Ethiopian sample sets; consistent with the results for nucleotide diversity.

6.1.5 Examining population differentiation at the CYP3A5 locus in Ethiopia

Inter-ethnic differentiation at the *CYP3A5* locus in Ethiopia was measured by calculating pairwise F_{ST} values and performing an exact test of population differentiation (both executed in Arlequin). Pairwise F_{ST} values were calculated using allele frequencies instead of haplotype information. This allowed for all polymorphic information, including singleton variants, to be accounted for when evaluating pairwise population differences. Pairwise F_{ST} values are shown in Table 6.6 and *p*-values from the exact test of population differentiation are shown in Table 6.7.

Table 6.6: Pairwise F_{ST} values, based on *CYP3A5* genotypic data for five Ethiopian populations. Pairwise F_{ST} values are shown in the bottom left side of the Table, the corresponding *p*-values are shown in the top right of the Table. *P*-values which are significant after Bonferroni correction (adjusted *p*-value = 0.0125; correction for 4 tests) are highlighted in green and the corresponding F_{ST} values are shown in red bold.

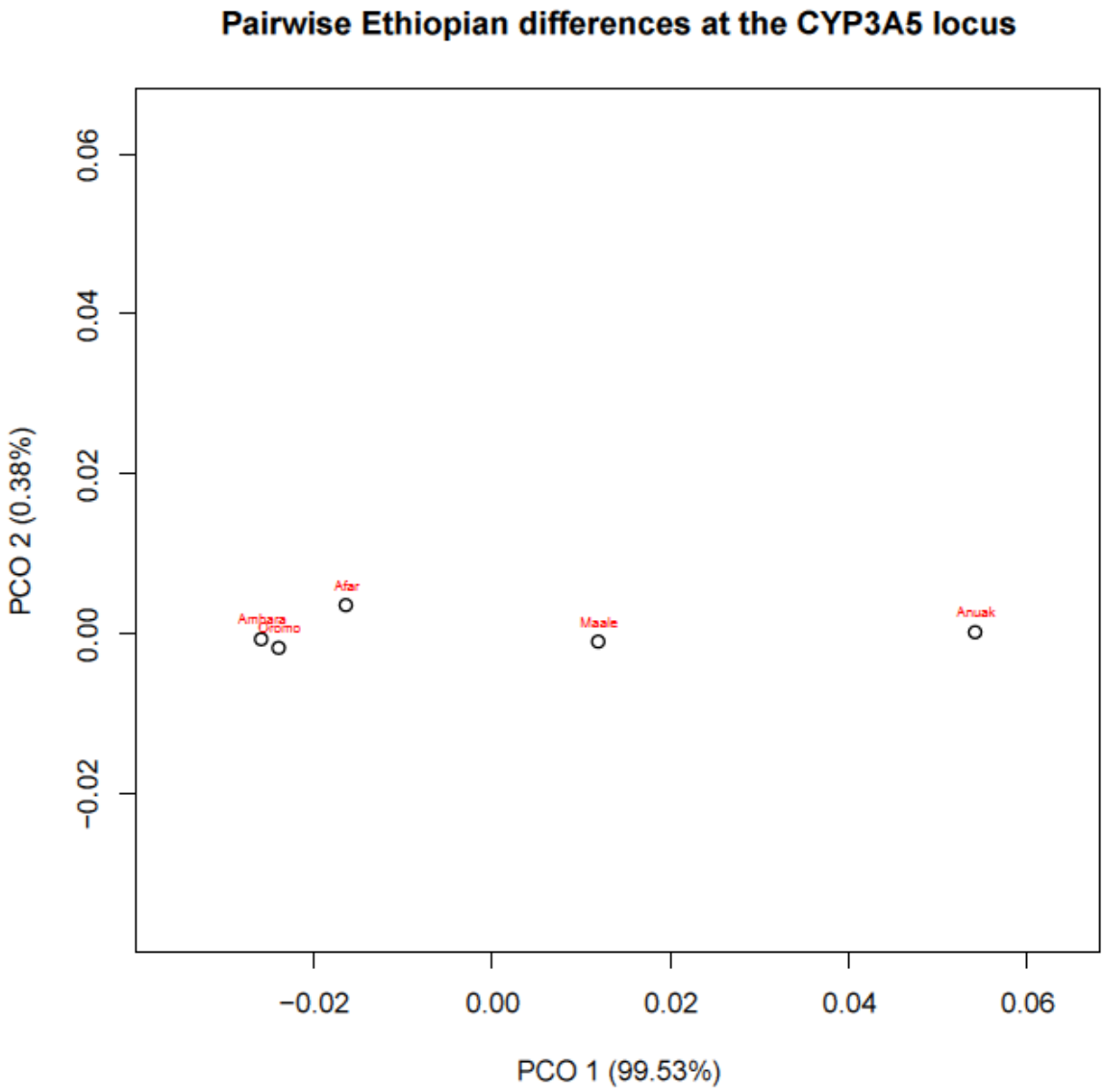
	Afar	Amhara	Anuak	Maale	Oromo
Afar	*	0.47708 ± 0.0047	**0.00000 ± 0.0000**	**0.00881 ± 0.0010**	0.56173 ± 0.0049
Amhara	-0.00102	*	**0.00000 ± 0.0000**	**0.00703 ± 0.0008**	0.91238 ± 0.0028
Anuak	**0.0801**	**0.091**	*	**0.00208 ± 0.0004**	**0.00000 ± 0.0000**
Maale	**0.01931**	**0.0212**	**0.02614**	*	**0.00743 ± 0.0009**
Oromo	-0.00205	-0.00375	**0.08902**	**0.02009**	*

Table 6.7: The resultant p-values from an exact test of population differentiation, based on Ethiopian genotypic data. Values shown in the Table correspond to *p*-values from pairwise population comparisons. *P*-values which are significant after Bonferroni correction (adjusted *p*-value = 0.0125; correction for 4 tests) are highlighted in green.

	Afar	Amhara	Anuak	Maale	Oromo
Afar	*				
Amhara	0.75305	*			
Anuak	**0.00000**	**0.00368**	*		
Maale	**0.00000**	0.21842	0.01469	*	
Oromo	0.60824	0.95138	**0.00126**	**0.00241**	*

A principal co-ordinates (PCO) plot was constructed, using pairwise F_{ST} values, to visualise population sub-structuring, see Figure 6.6. The Anuak significantly differ from the Afar, Amhara and Oromo; whereas the Maale are midway between Afro-Asiatic speakers and the Nilo-Saharan speaking Anuak.

Figure 6.6: A Principal Co-ordinates (PCO) plot based on pairwise F_{ST} values between the five Ethiopian populations. Pairwise F_{ST} values were calculated using Ethiopian *CYP3A5* genotype data for the 12,237bp re-sequenced region. Axis labels show the percentage of genetic distance captured by each axis.



6.2 Examining Ethiopian data in a global context

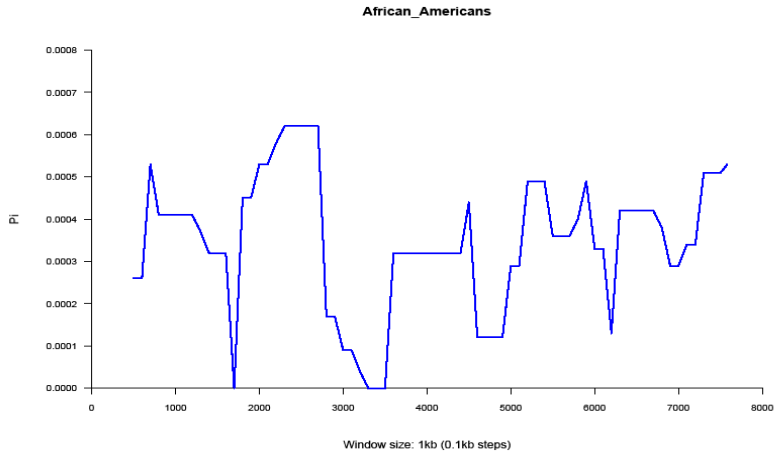
6.2.1 The global CYP3A5 allele frequency spectrum

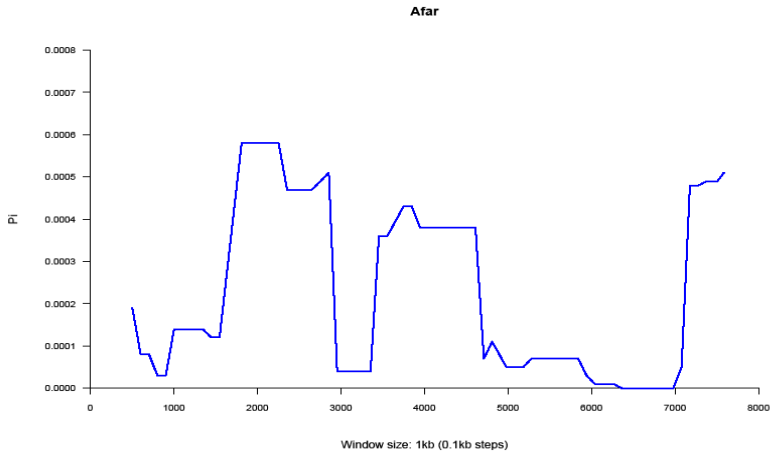
The African *CYP3A5* data were analysed in a global context by integrating the results with those previously published for datasets from the Coriell repositories (see chapter 2). A total of 8063bp of re-sequencing data, including the *CYP3A5* promoter and coding region, overlapped between the Ethiopian and Coriell datasets.

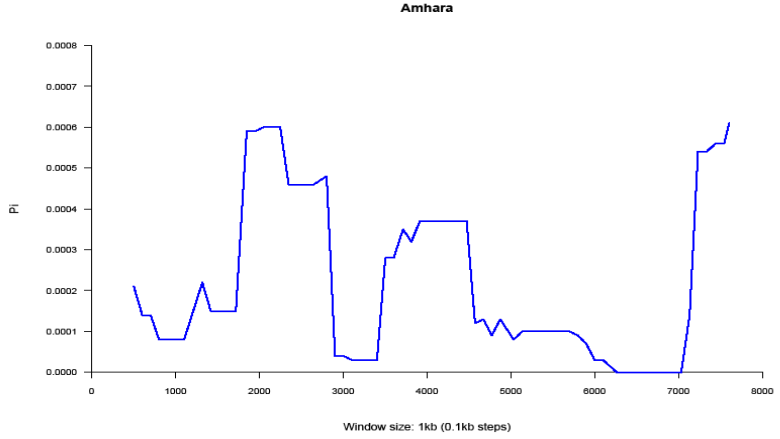
Nucleotide diversity is highest in African-Americans, the Anuak and Maale (Table 6.8). Nucleotide diversity levels are higher across the Ethiopian cohort than in Europeans and Han Chinese. Sliding windows analyses (using DnaSP software) were performed to analyse how nucleotide diversity varies across the re-sequenced 8063bp region in each population. A "diversity window" is a specific sized region, of the total data, for which π is calculated. The 8063bp fragment was divided up into overlapping windows (each 1000bp in length) which were generated every 100bp of sequence; π was then calculated for each window. This generated a series of π values which were plotted to show a "continuous" change in diversity, see Figure 6.7 (executed in the R-programming environment using a code written by Mr Pawel Zmarz).

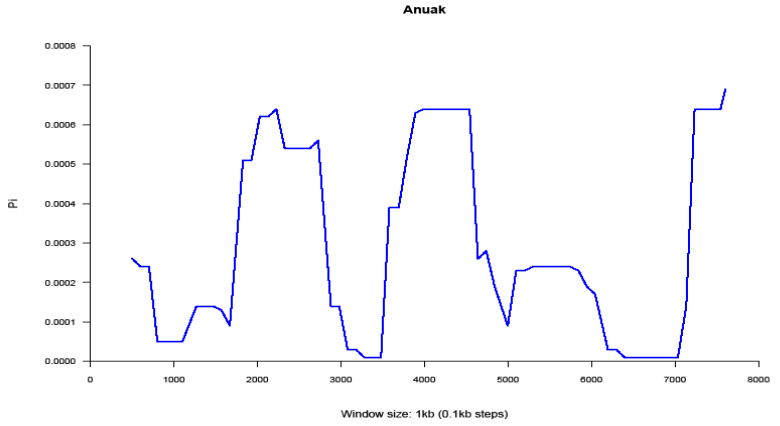
African-Americans appear to have more diversity across the region than all other global populations. Within each population, there were sharp rises in diversity within a region at position ~2200-2300 (corresponding to the *CYP3A5*1/CYP3A5*3* locus). Within all populations of recent African ancestry there was also a peak at position 4000, corresponding to the *CYP3A5*6* locus. *CYP3A5*7* was observed at low frequency in the global cohort (2 heterozygous individuals) and so did not affect diversity scores. Every population also had a peak at position at ~7700, corresponding to the position of the variant rs15524. This variant is often in high LD with the *CYP3A5*3* variant; although has no effect on gene transcription, mRNA processing or protein translation (Busi and Cresteil 2005).

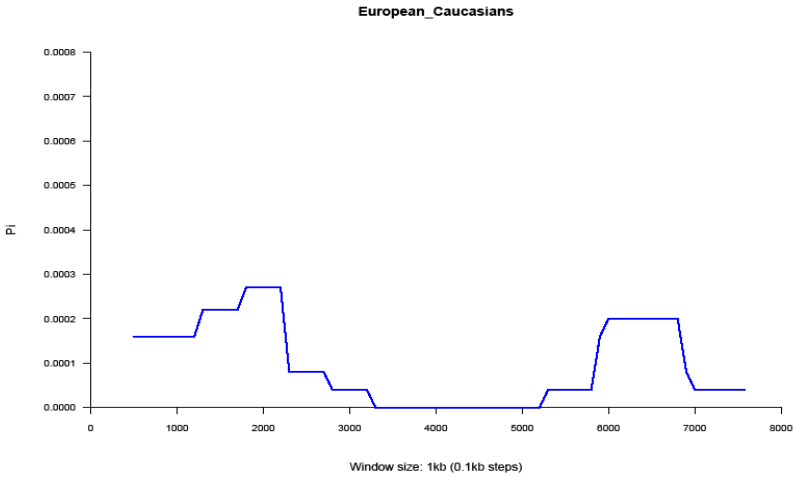
Figure 6.7: Sliding window analysis of diversity across an 8063bp *CYP3A5* region in 8 populations. Increases in pi (π) at position 2000 and 4000 correspond to *CYP3A5*1/CYP3A5*3* and *CYP3A5*6* loci; *CYP3A5*7* is observed at low frequency in the global cohort and so does not affect π values significantly. Peaks at position 7700 correspond to rs15524.

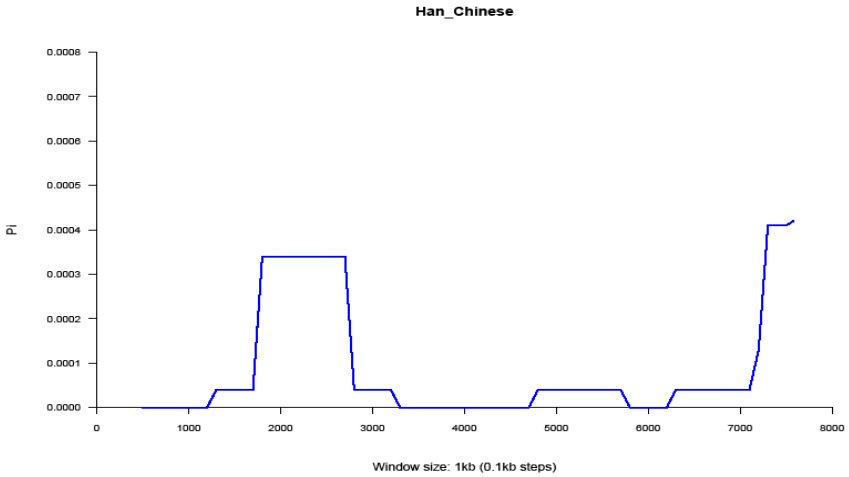


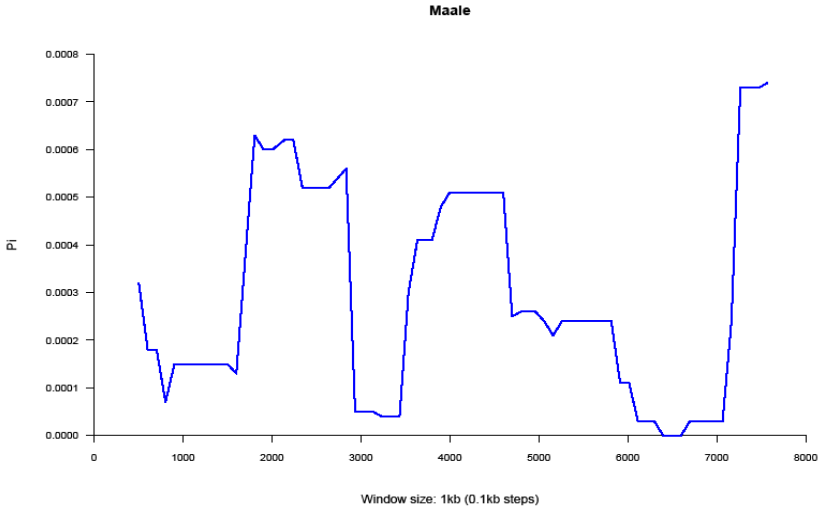


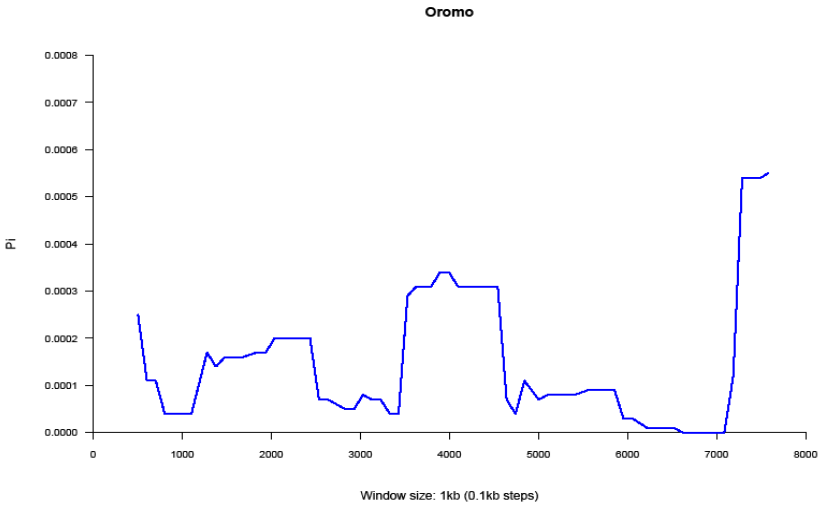





European_Caucasians
Pi
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
1000
2000
3000
4000
5000
6000
7000
8000
Window size: 1kb (0.1kb steps)


Han_Chinese
Pi
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
1000
2000
3000
4000
5000
6000
7000
8000
Window size: 1kb (0.1kb steps)


Maale
Pi
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
1000
2000
3000
4000
5000
6000
7000
8000
Window size: 1kb (0.1kb steps)


Oromo
Pi
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
1000
2000
3000
4000
5000
6000
7000
8000
Window size: 1kb (0.1kb steps)

6.2.2 Tests for departures from neutrality

Molecular diversity indices and tests of departures from neutrality are shown in (Table 6.8). A comparison of the ratio of synonymous to non-synonymous variation within human populations and between species (chimpanzee and human comparisons), by a McDonald-Kreitman test, was not significant. The results of the HKA test, comparing intra- and inter-species *CYP3A5* diversity were not significant (p=0.6346). A non-significant result suggests that the observed intra- and inter-species diversity do not significantly differ from neutral expectations (Kimura 1979). A non-significant p-value obtained here may be due to the paucity of variation, and excess of rare variants, identified in the global cohort.

The results of Tajima's *D,* Fu and Li's *D** and *F*,* Fu and Li's *F* and *D* (comparison with the chimpanzee), and Fu's *FS* analyses indicated a global skew towards rare variants (Table 6.8). Fu and Li's *D** and *F** analyses reported a significant departure from neutrality for both Europeans and the Anuak; although the results were only significant for individuals of recent European ancestry following Bonferonni correction for multiple testing. Fu and Li's *FS* reported a significant departure from neutrality for 7 of the 8 populations after Bonferonni correction. Strobeck's *S* calculates the probability of obtaining an equal number, or fewer, haplotypes than observed based on the gene frequency distribution. The results for all 8 populations were significant (p<0.0001 for every comparison); indicating that the *CYP3A5* haplotype structures observed in the global cohort, given the gene frequency data are atypical from neutral expectations. That is the distribution of all identified variants on haplotypes is not consistent with neutral expectations. The results from Strobeck's *S* are consistent with Fu's *FS* which found that the number of haplotypes defined by rare mutations significantly differs from neutral expectations.

An excess of high frequency derived alleles may be seen soon after a selective sweep has completed or during an ongoing selective sweep. This can be assessed using the *H* test (Fay and Wu 2000). The results of the *H* test for *CYP3A5* were not statistically significant within any population (p>0.05), however nucleotide diversity is low at this locus and this may be affecting the calculation, as reported previously (Thompson et al. 2004).

Table 6.8: A summary of the tests for departures from neutrality for the 8063bp overlapping region of *CYP3A5*. Statistically significant departures from neutrality, following Bonferonni correction (correction for 8 tests; adjusted $p \leq 0.00625$) are shown in bold and highlighted in green.

	Global populations			Ethiopian populations				
	African-Americans	**Europeans**	**Han Chinese**	**Afar**	**Amhara**	**Anuak**	**Maale**	**Oromo**
Sample size	23	24	23	75	76	76	76	76
Nucleotide diversity (π)	5.4×10^{-4}	9×10^{-5}	1.1×10^{-4}	2.1×10^{-4}	2.5×10^{-4}	3.6×10^{-4}	3.5×10^{-4}	2.2×10^{-4}
McDonald-Kreitman test	0.475	0.50	0.777	0.462	0.475	0.576	1.00	0.777
Tajima's *D*	-1.04	-1.92	-1.21	-1.46	-1.13	-1.26	-0.96	-1.79
Fu and Li's *D**	-0.97	**-2.86**	-1.82	-1.19	0.12	-2.72	-0.11	-1.05
Fu and Li's *F**	-1.17	**-3.00**	-1.91	-1.54	-0.43	-2.57	-0.53	-1.58
Fu and Li's *D*	-0.64	-1.37	-1.45	-1.56	-0.08	-1.93	0.73	-0.96
Fu and Li's *F*	-0.92	-1.81	-1.55	-1.79	-0.52	-1.96	0.13	-1.48
Fu's *FS*	**-31.06**	**-5.71**	-1.54	**-9.48**	**-11.74**	**-18.44**	**-11.08**	**-22.84**
Strobeck's *S*	**1.00**	**0.999**	0.929	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Fay and Wu's *H* statistic	-0.13140	-3.25532	-1.89372	0.10774	-0.23998	-0.47177	-0.87086	-1.75514

N.B: Statistics highlighted in blue were performed using overlapping sequence from the chimpanzee genome as an out-group. For the McDonald-Kreitman test (based on coding region variation only) and Strobeck's *S* values are shown.

6.2.3 Haplotype associations of CYP3A5 alleles

Figures 6.8a-e show inferred haplotypes, for the 8063bp region, within each of the 8 populations. The inferred haplotypes confirm the results from sliding windows analyses; there are four main haplogroups, each defined by one of *CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*7*, although three low frequency *CYP3A5*3/*6* haplotypes were observed.

*CYP3A5*3, CYP3A5*6* and *CYP3A5*7* haplotypes are characterised by a paucity of additional variation. Fewer *CYP3A5*3* haplotypes were observed in Ethiopia than in the Coriell datasets; although *CYP3A5*3* haplotype diversity is higher in the Oromo and Amhara than in other Ethiopian groups. *CYP3A5*6* haplotypes are largely homogeneous between all populations in which they were observed. *CYP3A5*7* was only observed, at low frequency, in Ethiopia. Diversity was highest in the *CYP3A5*1* haplogroup.

Approximately 98% of European and 83% of Han Chinese haplotypes are defined by the *CYP3A5*3* mutation as are the majority of Ethiopian Afar (~64%), Amhara (~67%) and Oromo (~67%) chromosomes (Figure 5.8). Notably, the haplotypes observed in these three Ethiopian groups are typical of those observed in non-Africans. Within an 8063bp region of *CYP3A5*, fewer *CYP3A5*3* and *CYP3A5*6* haplotypes have been inferred from the genotype data than for a 12,237bp region in the same individuals. A comparison of inferred haplotypes for the two re-sequenced regions suggests there are more high frequency derived variants in *CYP3A5* flanking regions, as little as ≤2000bp in size, than across the actual gene region itself (see section 5.2.5). These can define additional *CYP3A5*3* and *CYP3A5*6* haplotypes, although are unlikely to affect *CYP3A5* transcription or translation.

Figure 6.8a: Inferred *CYP3A5*3* haplotypes for 8 populations. Haplotype information for African-American, European and Han Chinese populations has been adapted from (Thompson et al. 2004), so that only data for the overlapping genomic region are shown. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The locus which defines the *CYP3A5*1/CYP3A5*3* alleles are highlighted in green. "***N***" refers to the number of chromosomes that are of a particular haplotype.

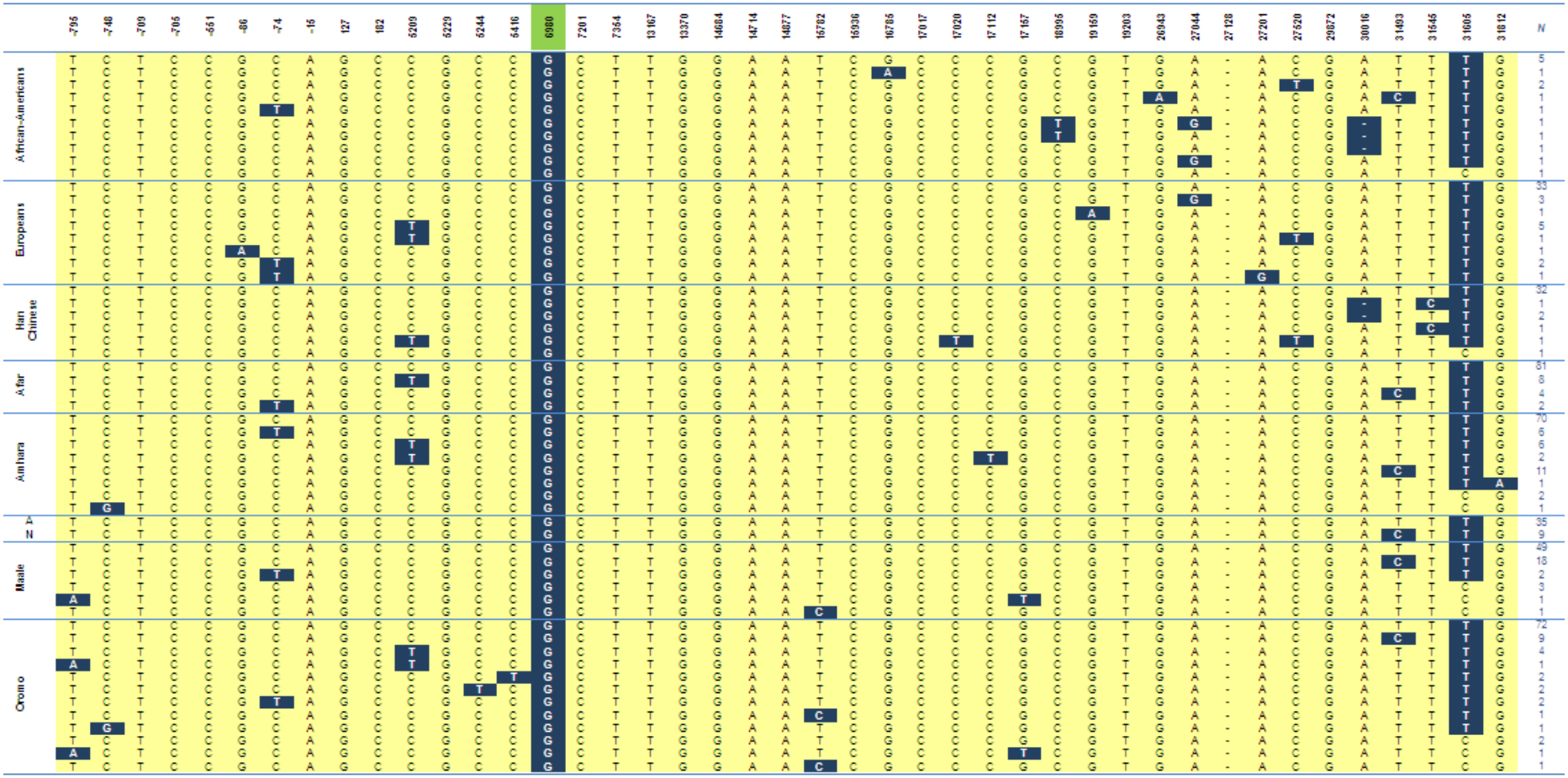


Figure 6.8b: Inferred *CYP3A5*6* haplotypes for 8 populations. Haplotype information for African-American, European and Han Chinese populations has been adapted from (Thompson et al. 2004), so that only data for the overlapping genomic region are shown. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The locus which defines the *CYP3A5*6* allele is highlighted in green. "***N***" refers to the number of chromosomes that are of a particular haplotype.

***Population codes:** AA: African-Americans; AF: Afar; AM: Amhara; AN: Anuak; ML: Maale; OR: Oromo

	-795	-748	-709	-705	-551	-86	-74	-15	127	182	5209	5229	5244	5416	6980	7201	7354	13167	13370	14684	14714	14877	15782	15936	16785	17017	17020	17112	17157	18995	19159	19203	26943	27044	27128	27201	27520	29872	30016	31493	31545	31605	31812	*N*
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	3
AA	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	**T**	G	A	T	T	C	G	1
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	**-**	T	T	C	G	2
AF	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	26
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	**C**	G	A	-	A	C	G	A	T	T	C	G	1
AM	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	20
AN	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	35
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	**T**	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	1
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	**T**	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	2
	T	C	T	C	**A**	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	**-**	T	T	**T**	G	1
ML	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	20
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	**C**	G	A	-	A	C	G	A	T	T	C	G	2
OR	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	**A**	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	18

Figure 6.8c: Inferred *CYP3A5*7* haplotypes for 8 populations. Haplotype information for African-American, European and Han Chinese populations has been adapted from (Thompson et al. 2004), so that only data for the overlapping genomic region are shown. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The locus which defines the *CYP3A5*7* allele is highlighted in green. "***N***" refers to the number of chromosomes that are of a particular haplotype.

***Population codes:** AN: Anuak; ML: Maale

	-795	-748	-709	-705	-551	-86	-74	-15	127	182	5209	5229	5244	5416	6980	7201	7354	13167	13370	14684	14714	14877	15782	15936	16785	17017	17020	17112	17157	18995	19159	19203	26943	27044	27128	27201	27520	29872	30016	31493	31545	31605	31812	*N*
AN	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	T	A	C	G	A	T	T	C	G	1
ML	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	T	A	C	G	A	T	T	C	G	1

Figure 6.8d: Inferred recombinant *CYP3A5*3/*6* haplotypes for 8 populations. Haplotype information for African-American, European and Han Chinese populations has been adapted from (Thompson et al. 2004), so that only data for the overlapping genomic region are shown. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. Each row corresponds to a haplotype and each column to a polymorphic site. The numbers above each column correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). The two loci which define the *CYP3A5*3* and *CYP3A5*6* alleles are highlighted in green. "***N***" refers to the number of chromosomes that are of a particular haplotype.

***Population codes:** AM: Amhara; ML: Maale; OR: Oromo

	-795	-748	-709	-705	-551	-86	-74	-15	127	182	5209	5229	5244	5416	6980	7201	7354	13167	13370	14684	14714	14877	15782	15936	16785	17017	17020	17112	17157	18995	19159	19203	26943	27044	27128	27201	27520	29872	30016	31493	31545	31605	31812	*N*
AM	T	C	T	C	C	G	C	A	G	C	C	G	C	C	G	C	T	T	G	A	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	2
	T	C	T	C	C	G	C	A	G	C	C	G	C	C	G	C	T	T	G	A	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	T	G	1
ML	T	C	T	C	C	G	C	A	G	C	C	G	C	C	G	C	T	T	G	A	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	1

Figure 6.8e: Inferred *CYP3A5*1* haplotypes for 8 populations. Haplotype information for African-American, European and Han Chinese populations has been adapted from (Thompson et al. 2004), so that only data for the overlapping genomic region are shown. Positions marked in yellow are ancestral alleles at each position (as inferred from the chimpanzee sequence) and blue are the derived. The numbers above each polymorphic site correspond to the position of each polymorphic site relative to the ATG start codon (NCBI, Build 132, http://www.ncbi.nlm.nih.gov/). "*N*" refers to the number of chromosomes that are of a particular haplotype.

***Population codes:** AS: Han Chinese; EU: Europeans

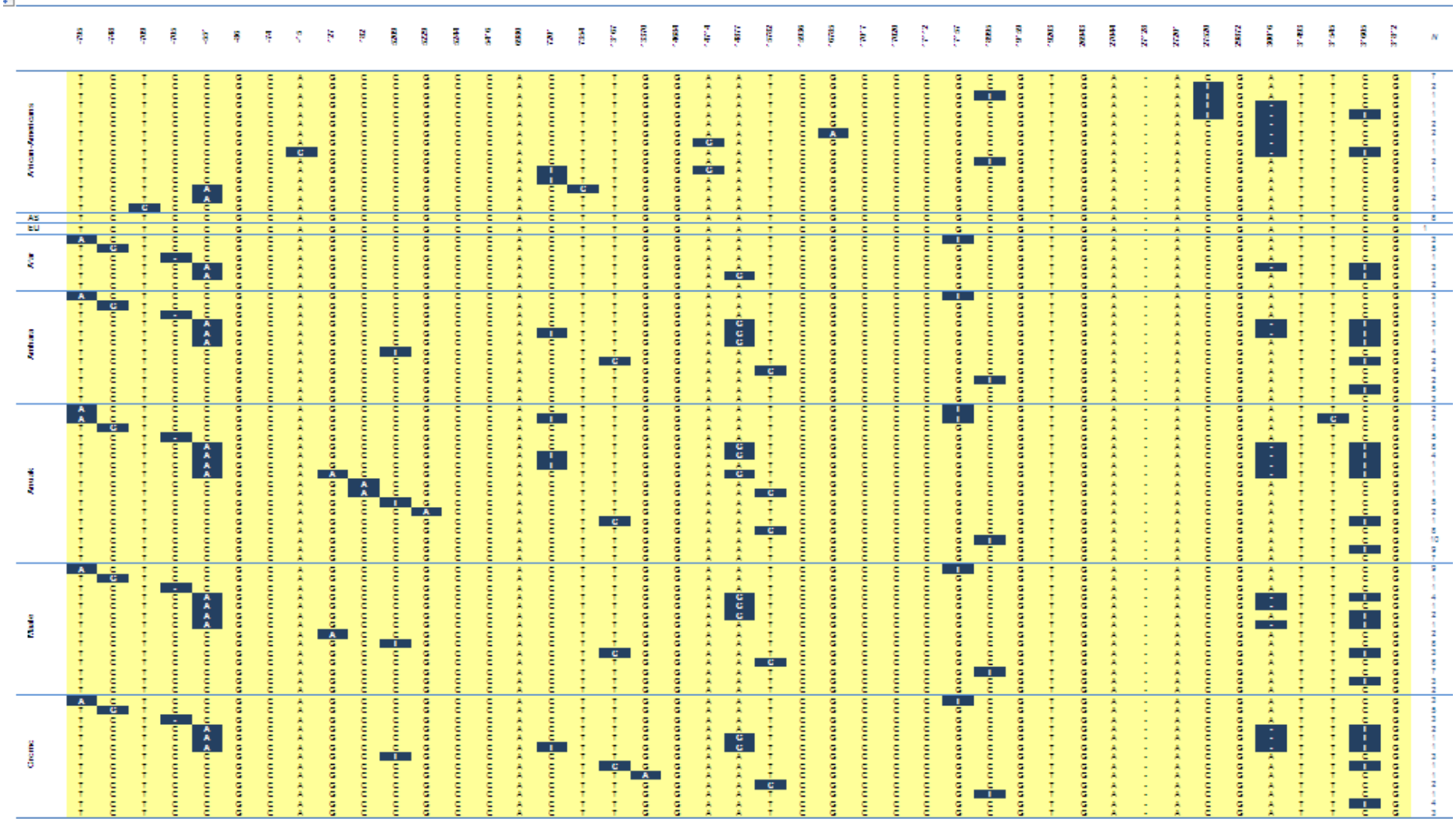


Figure 6.9a: The proportion of inferred *CYP3A5* haplotypes, for an 8063bp region, in each of 8 populations. Each haplotype has been numbered from 1-72 and numbers in parentheses indicate the associated haplogroup (*CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*3/*6* or *CYP3A5*7*).

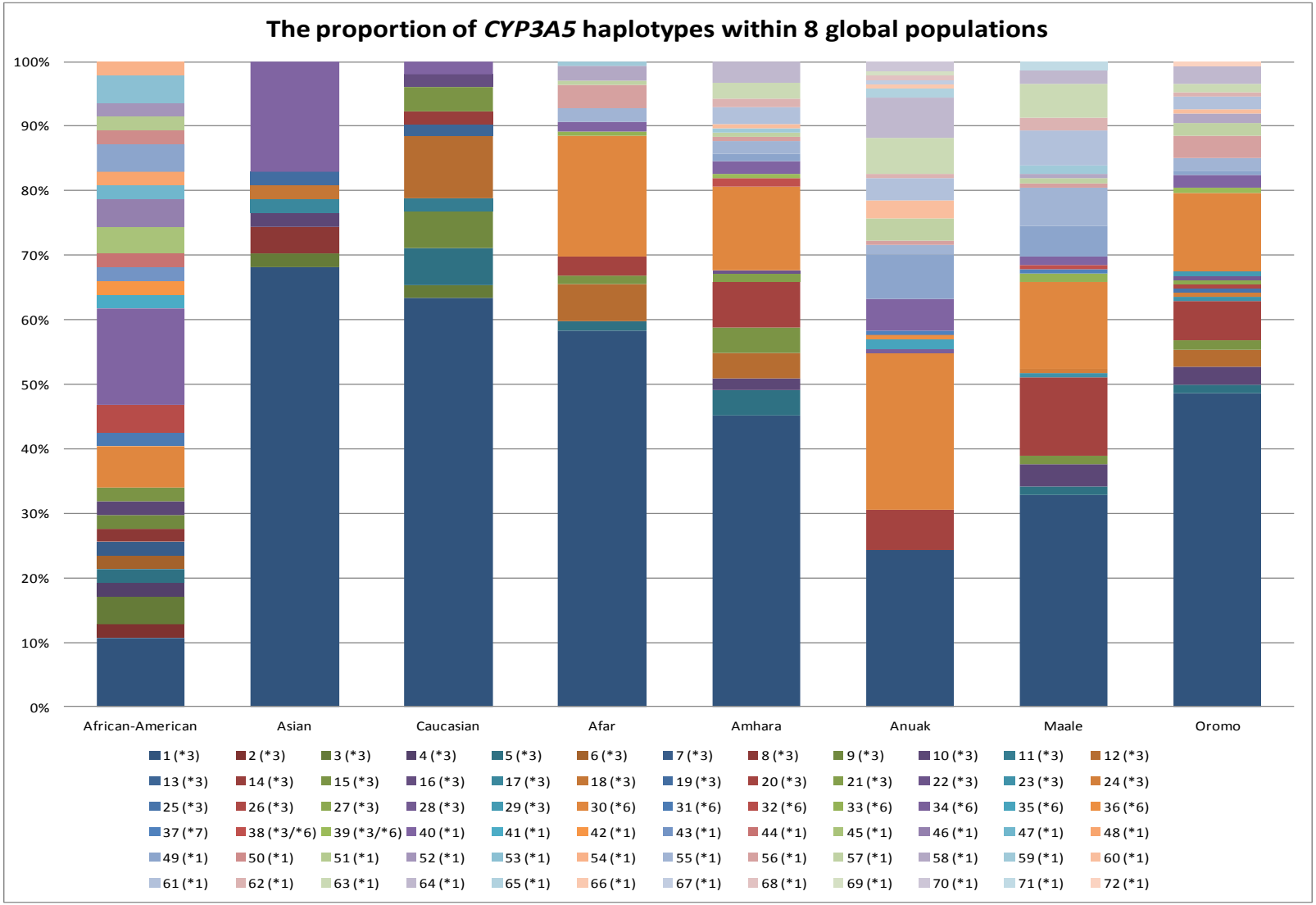


Figure 6.9b: Global diversity in the *CYP3A5*1* haplogroup

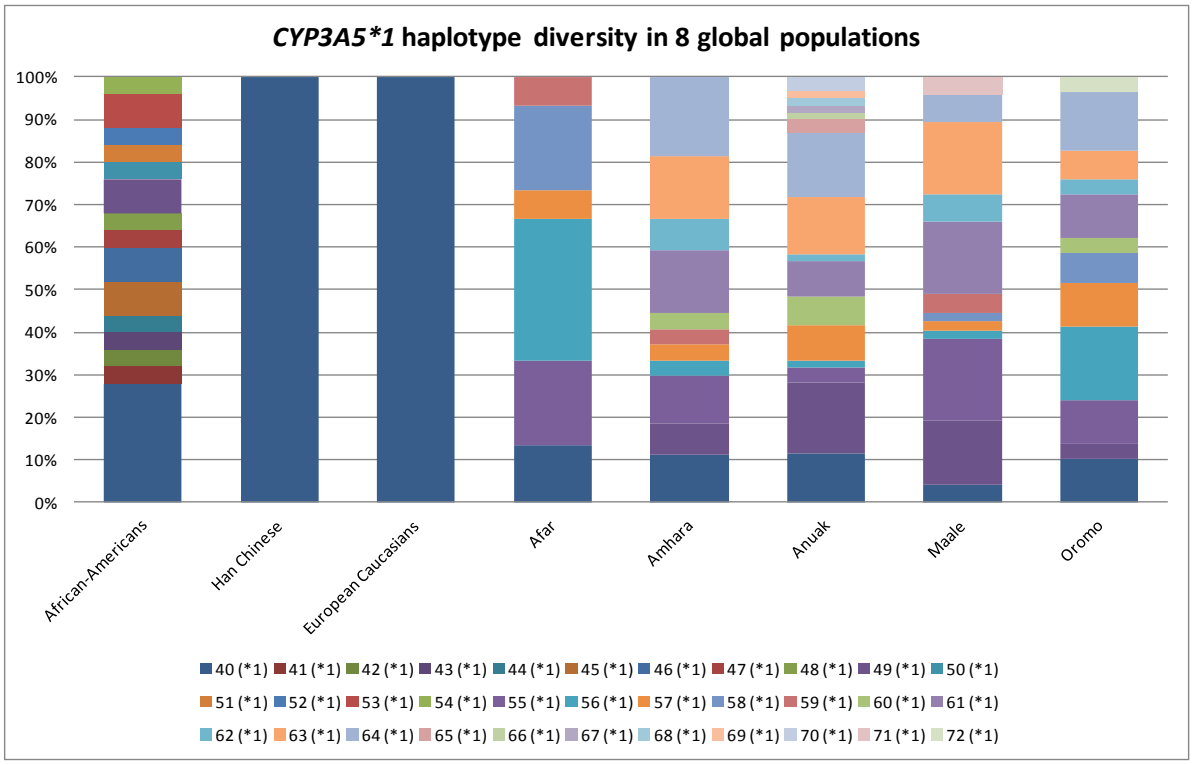


Figure 6.9c: Global diversity in the *CYP3A5*3* haplogroup

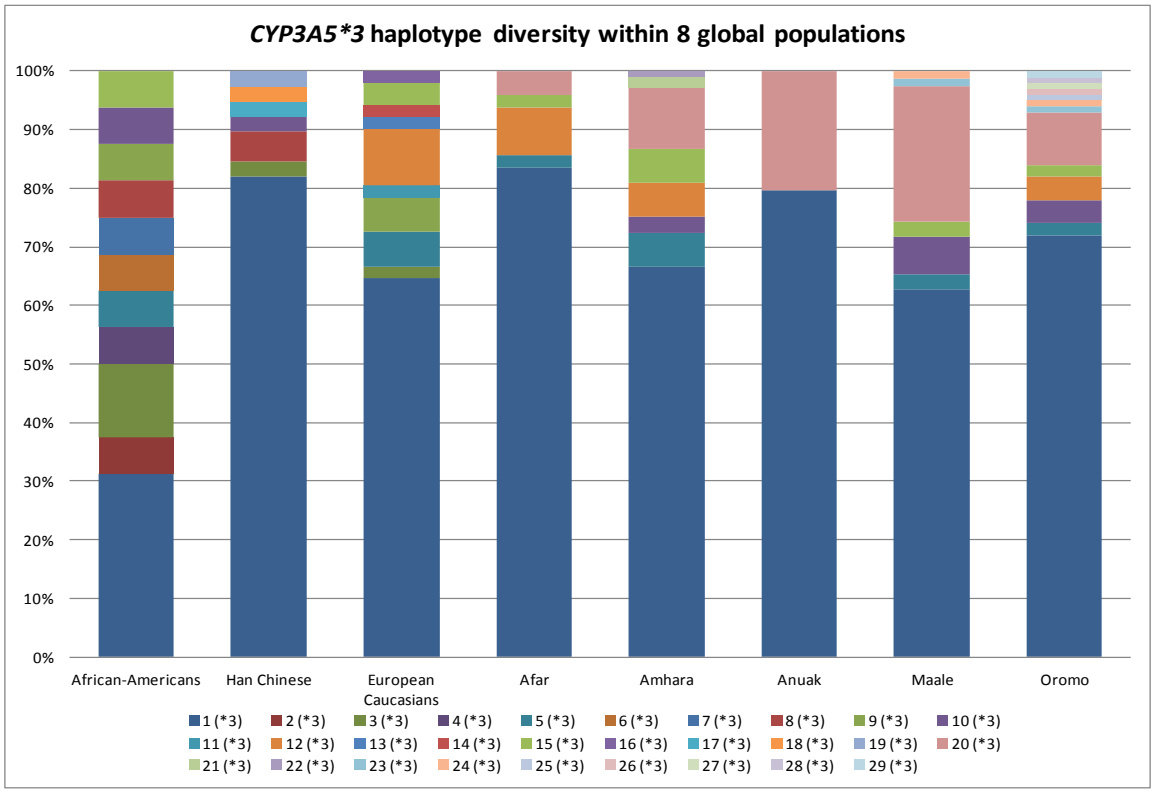


Figure 6.9d: Global diversity in the *CYP3A5*6* haplogroup

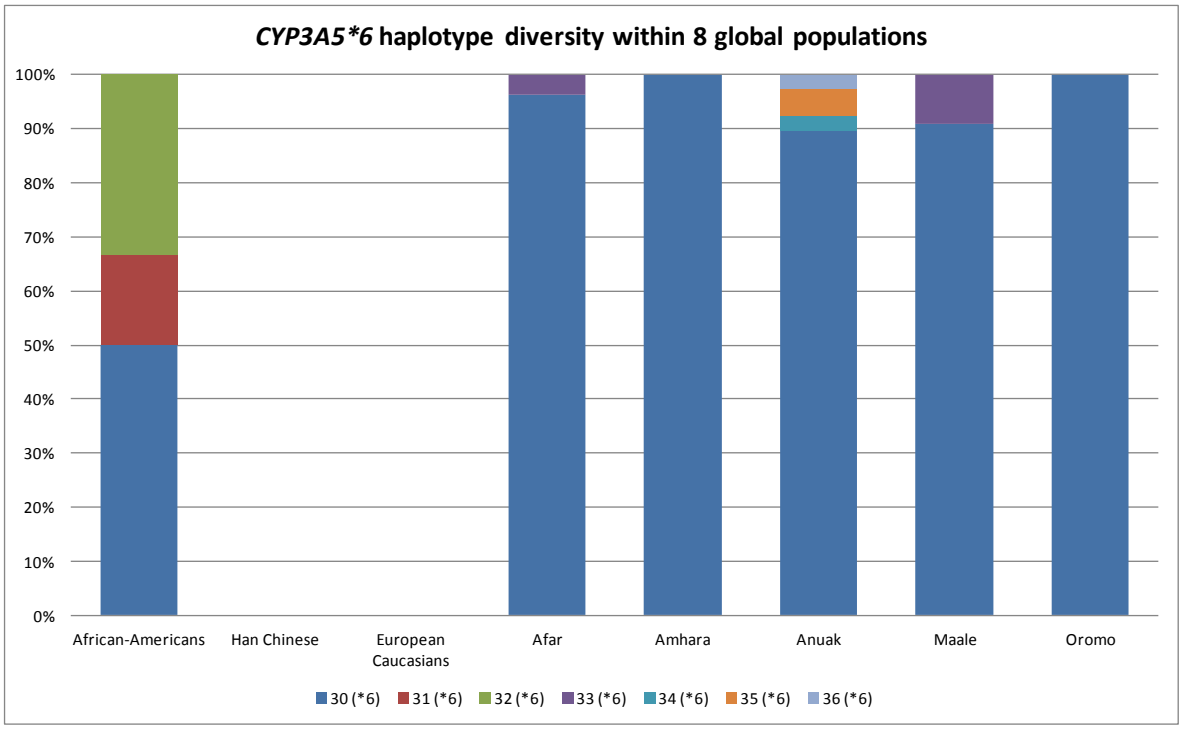


Figure 6.9e: Global diversity in the *CYP3A5*7* haplogroup

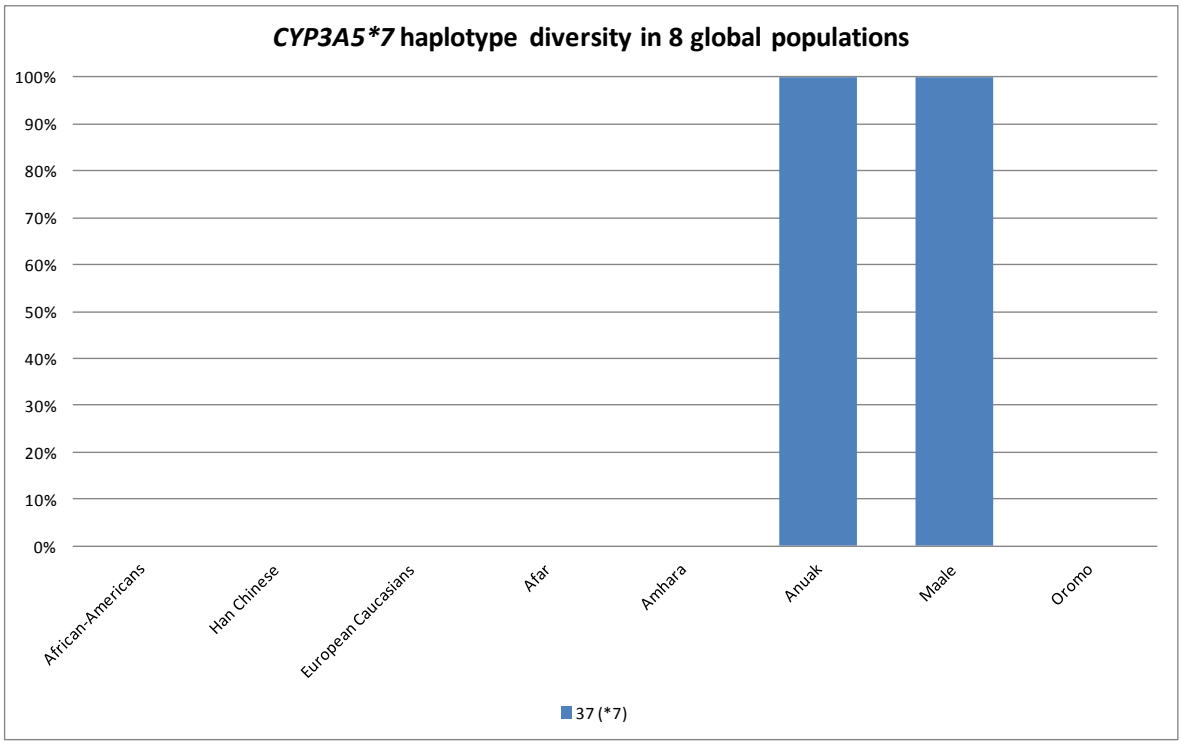
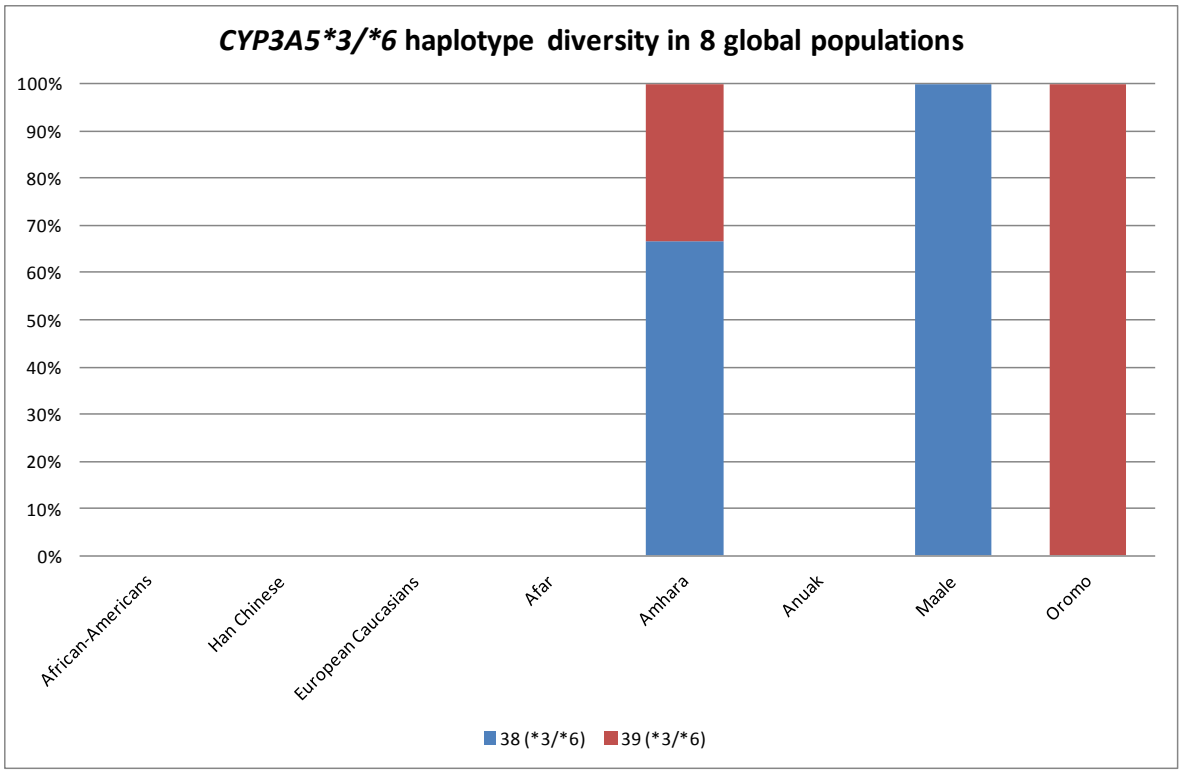


Figure 6.9f: Global diversity in the *CYP3A5*3/*6* haplogroup



Note that *CYP3A5*6, CYP3A5*7* and *CYP3A5*3/*6* haplotypes are not observed in every global population (see Figures 5.9d-f).

6.2.4 Analysing Ethiopian haplotype diversity in a global context

Gene diversity appears to be highest in African-Americans and lowest in Europeans and Han Chinese (Figure 6.9a). The *CYP3A5*1* haplogroup is most diverse in the Anuak and Maale (Figure 6.9b), but the Anuak have the lowest levels of *CYP3A5*3* haplogroup diversity in Ethiopia (Figure 6.9c); the Amhara and the Oromo have the highest. African-Americans also have high levels of haplotype diversity; which appears to be higher than the Ethiopian groups (Figures 6.9a-d).

A global comparison of haplotype diversity, see Table 6.9, found that all five Ethiopian groups significantly differ from other global populations. Inter-Ethiopian comparisons were consistent with all results reported throughout this thesis; the Afar, Amhara and Oromo cluster together whereas the Anuak differ from every Ethiopian population. The Maale are intermediate between the Afar, Amhara and Oromo cluster and the Anuak. Similarities between the Maale, Amhara and Oromo may be due to their similar frequencies and compositions of *CYP3A5*3* haplotypes. African-Americans differ from all other global populations, as do Europeans and Han Chinese.

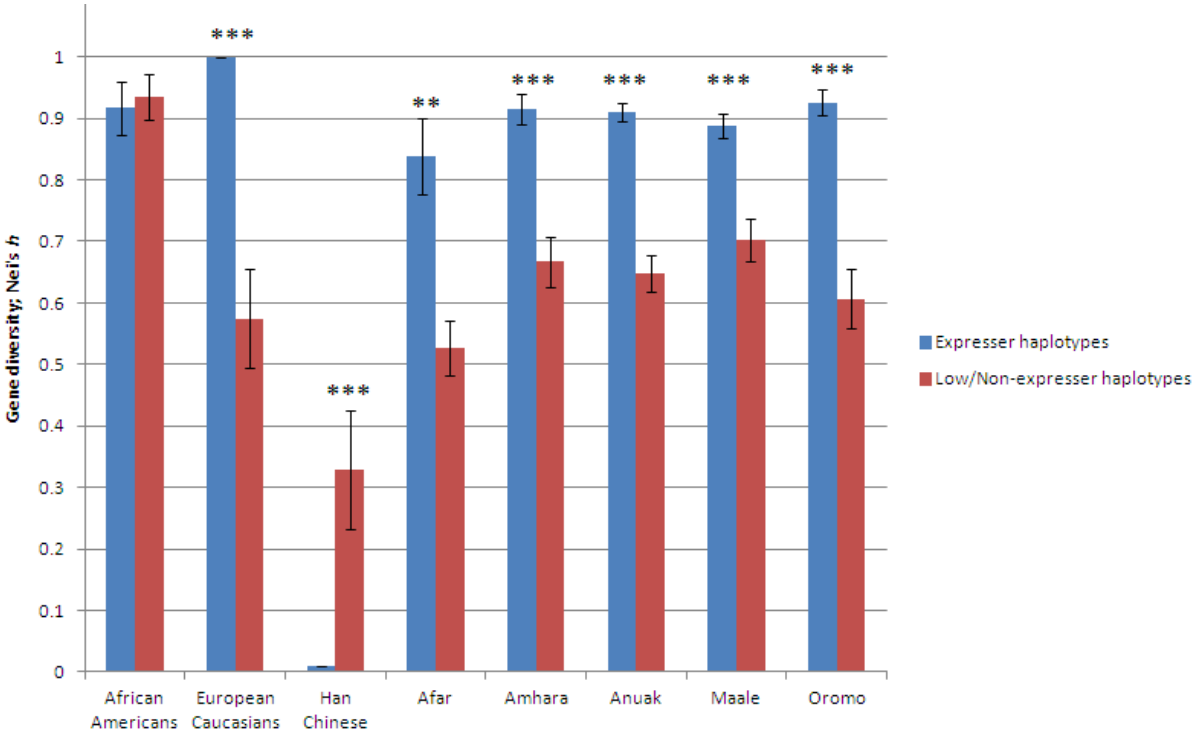
Table 6.9: Nei's h to compare heterozygosity in an 8063bp *CYP3A5* in 8 populations. Nei's *h* was calculated using the *h*-diff test (Thomas et al. 2002) and implemented in the R-programming environment. Significant pairwise differences were assessed using an exact test of population differentiation (executed in Arlequin software). Pairwise comparisons which were significant after Bonferroni correction are highlighted in green [adjusted *p*-value = 0.00625; correction for 8 tests].

Nei's *h*	0.963	0.514	0.589	0.623	0.770	0.865	0.850	0.744
	African-Americans	**Han Chinese**	**Europeans**	**Afar**	**Amhara**	**Anuak**	**Maale**	**Oromo**
African-Americans	*							
Han Chinese	**<0.00001**	*						
Europeans	**<0.00001**	**0.00046**	*					
Afar	**<0.00001**	**<0.00001**	**0.00007**	*				
Amhara	**<0.00001**	**<0.00001**	**0.00056**	0.002	*			
Anuak	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	*		
Maale	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	0.032	**0.00026**	*	
Oromo	**<0.00001**	**0.00181**	**0.00040**	0.129	0.791	**<0.00001**	0.011	*

Table 6.10: Nei's h to compare heterozygosity in each of five major *CYP3A5* haplotypes classes in eight populations. Nei's *h* was calculated using the *h*-diff test (Thomas et al. 2002) and implemented in the R-programming environment. Pairwise comparisons which were significant after Bonferroni correction are highlighted in green [adjusted *p*-value = 0.01; correction for 5 tests].

Nei's *h*	0.921	0.481	0.146	0.000	0.6
	CYP3A5*1	***CYP3A5*3***	***CYP3A5*6***	***CYP3A5*7***	***CYP3A5*3/CYP3A5*6***
CYP3A5*1	*				
CYP3A5*3	<0.000001	*			
CYP3A5*6	<0.000001	<0.000001	*		
CYP3A5*7	<0.000001	<0.000001	0.0005	*	
CYP3A5*3/CYP3A5*6	0.068	0.849	0.172	0.179	*

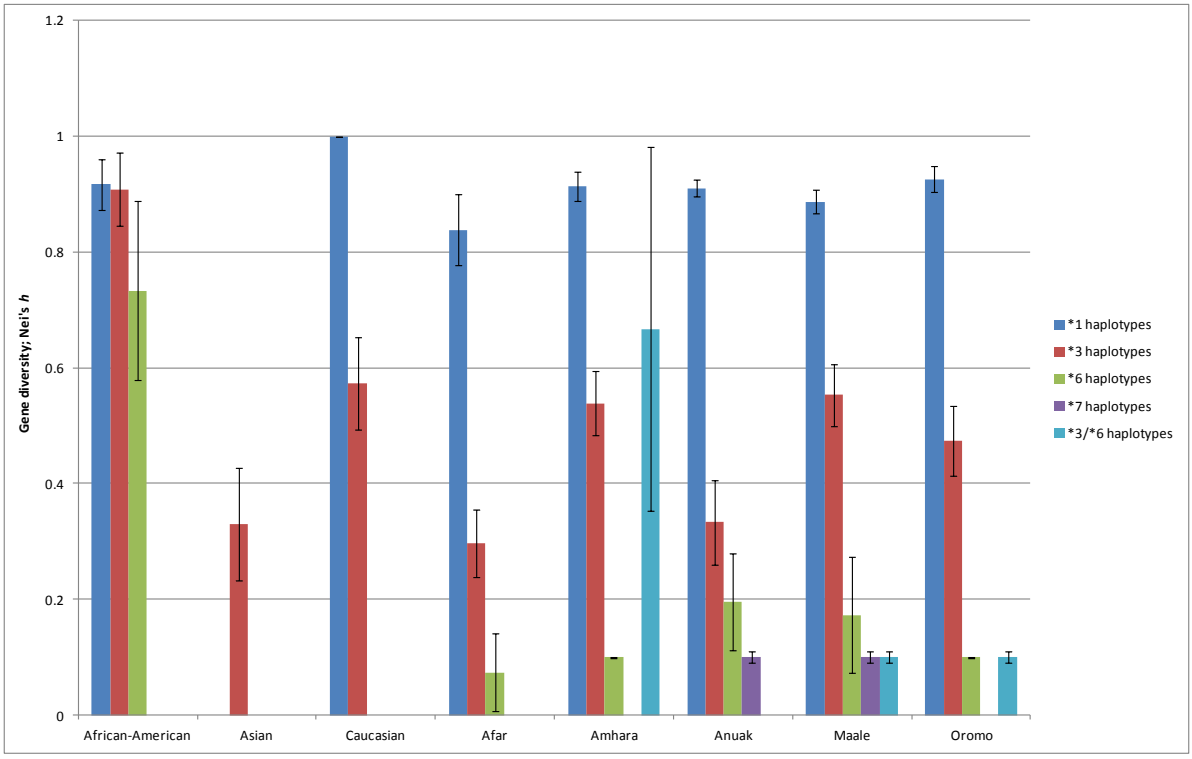
Figure 6.10: Pairwise intra-populations comparisons of expresser and low/non-expresser *CYP3A5* haplotype heterozygosity. Nei's *h* and significance of differences were measured using *h*-diff (Thomas et al. 2002). *** indicates comparisons which were significant following Bonferroni correction (for 8 tests; adjusted *p*-value=0.00625), ** indicates a comparison was significant before, but not after, Bonferonni correction for multiple testing.



A comparison of expresser and low/non-expresser haplotypes found that expresser haplotypes are significantly more diverse in every population except African-Americans and the Afar.

The *CYP3A5*1* haplogroup is significantly more diverse than the other haplotype classes (p<0.0001 for every comparison), and had over twice as much diversity than the *CYP3A5*3* haplogroup (Figure 5.11). *CYP3A5*3* haplotypes are significantly more diverse than *CYP3A5*6* and *CYP3A5*7*; although the latter are observed at a lower global frequency.

Figure 6.11: Comparisons of gene diversity in each haplotype class by global population. Nei's *h* was calculated to estimate gene diversity using the *h*-diff program (Thomas et al. 2002). Error bars denote standard deviation.



A comparison of diversity within expresser and within low/non-expresser haplotypes between populations found that African-Americans significantly differed from all other populations; and Han Chinese from three of the five Ethiopian groups. Interestingly, no significant differences, in diversity levels of low/non-expresser haplotypes, were observed between Europeans and Ethiopian populations. Levels of gene diversity in low/non-expresser haplotypes were highest in the Maale. Within the remaining populations gene diversity was significantly higher in expresser haplotypes than in low/non-expresser haplotypes (Table 6.11).

Table 6.11a: Pairwise comparisons of heterozygosity in expresser haplogroups by an exact test of population differentiation. *P* values which are significant after Bonferroni correction are highlighted in green [adjusted *p*-value = 0.00625; correction for 8 tests].

Nei's *h*	0.917	0.000	0.000	0.838	0.915	0.910	0.887	0.926
	African-Americans	Han Chinese	European	Afar	Amhara	Anuak	Maale	Oromo
African-Americans	*							
Han Chinese	**<0.0000001**	*						
European	**<0.0000001**	1.000000	*					
Afar	0.297	**<0.0000001**	**<0.0000001**	*				
Amhara	0.966	**<0.0000001**	**<0.0000001**	0.248	*			
Anuak	0.888	**<0.0000001**	**<0.0000001**	0.252	0.882	*		
Maale	0.713	**<0.0000001**	**<0.0000001**	0.447	0.680	0.357	*	
Oromo	0.848	**<0.0000001**	**<0.0000001**	0.175	0.728	0.985	0.359	*

Table 6.11b: Pairwise comparisons of heterozygosity in low/non-expresser haplogroups, by an exact test of population differentiation. *P* values which are significant after Bonferroni correction are highlighted in green [adjusted *p*-value = 0.00625; correction for 8 tests].

Nei's *h*	0.935	0.329	0.573	0.528	0.667	0.648	0.703	0.607
	African-Americans	Han Chinese	European	Afar	Amhara	Anuak	Maale	Oromo
African-Americans	*							
Han Chinese	**4.97×10^{-9}**	*						
European	**2×10^{-4}**	0.05	*					
Afar	**1.38×10^{-12}**	0.06	0.658	*				
Amhara	**8×10^{-4}**	**0.001**	0.294	0.02	*			
Anuak	**2×10^{-4}**	**0.002**	0.373	0.02	0.712	*		
Maale	**0.0018**	**0.0003**	0.136	**0.003**	0.531	0.234	*	
Oromo	**4.17×10^{-8}**	0.01	0.71	0.225	0.344	0.5	0.114	*

6.2.5 Population differentiation between Ethiopians and other global populations

Population differentiation at the *CYP3A5* locus was measured by pairwise F_{ST} comparisons and by an exact test of population differentiation. Allelic data were used to evaluate population differentiation, allowing for all polymorphic information, including singleton variants, to be accounted for. Pairwise F_{ST} results are shown in Table 6.12 and the results from the exact test of population differentiation are shown in Table 6.13. Pairwise F_{ST} results were used to produce principal co-ordinates (PCO) plots to visually represent genetic similarities and differences, see Figure 6.12.

The results from all three analyses found that the Ethiopian groups split into two distinct clusters (as seen in Figure 6.6). The Afar, Amhara and Oromo are intermediate between the Han Chinese individuals and other Ethiopian groups. African-Americans are much more similar to the Anuak than to other populations with recent African ancestry (Figure 6.12). These results are an extension of those presented in chapters 3 and 4, in which

the Afar, Amhara and Oromo are characteristic of non-African populations at the *CYP3A5* locus.

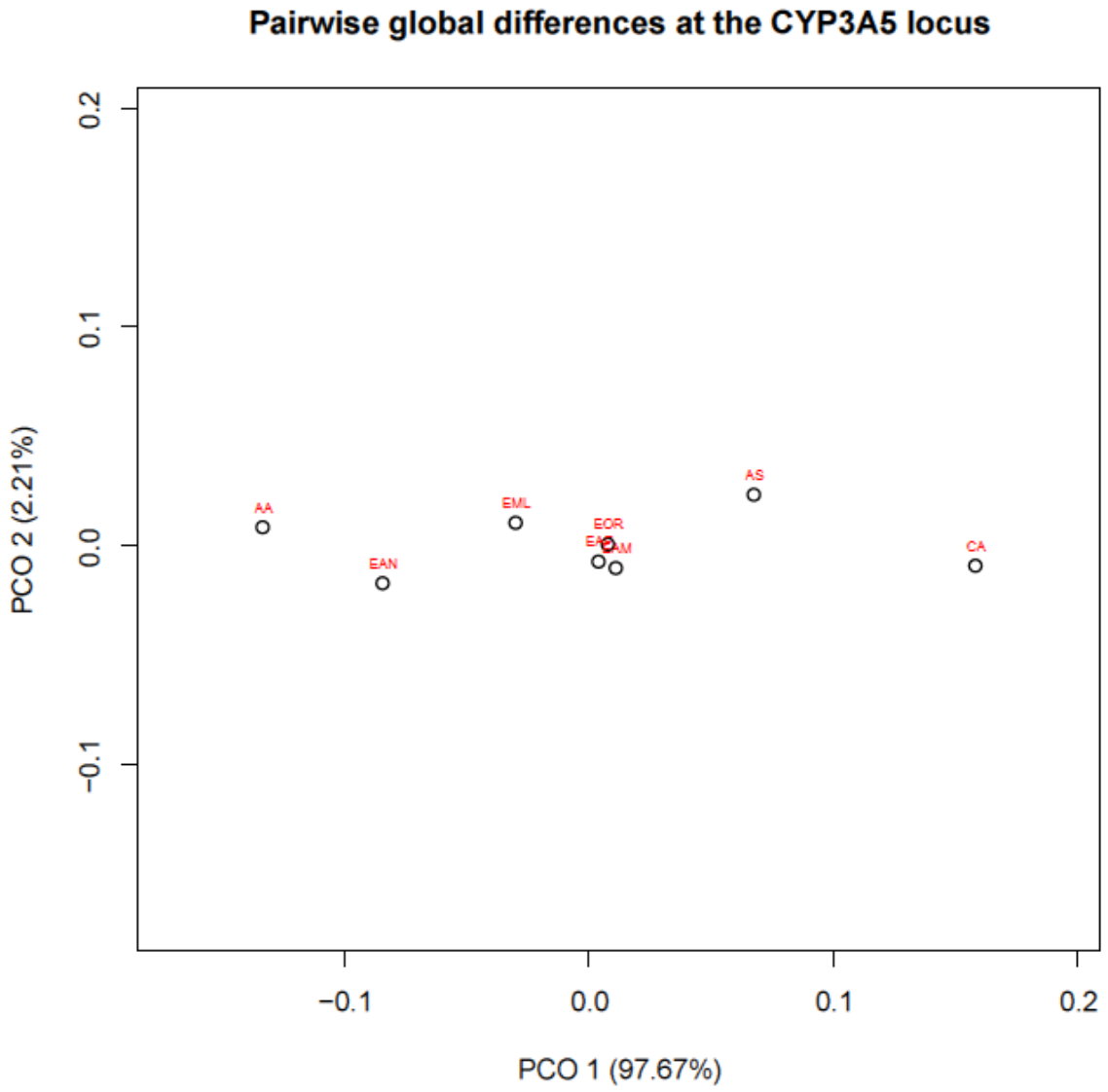
Table 6.12: Pairwise F_{ST} values based on overlapping *CYP3A5* genotypic data for five Ethiopian populations and three other global populations. Pairwise F_{ST} values are shown in the bottom left side of the Table, the corresponding *p*-values are shown in the top right of the Table. *P*-values which are significant after Bonferroni correction (adjusted p-value = 0.00625; correction for 8 tests) are highlighted in green.

	Afar	Amhara	Anuak	Maale	Oromo	African-Americans	Europeans	Han Chinese
Afar	*	0.74597	<0.00001	0.00436	0.76537	<0.00001	<0.00001	<0.00001
Amhara	-0.00248	*	<0.00001	0.00347	0.93525	<0.00001	<0.00001	<0.00001
Anuak	**0.04566**	**0.05138**	*	0.00257	<0.00001	<0.00001	<0.00001	<0.00001
Maale	**0.01951**	**0.01736**	0.01061	*	0.00267	<0.00001	<0.00001	<0.00001
Oromo	-0.00255	-0.0036	**0.04981**	0.01547	*	<0.00001	<0.00001	<0.00001
African-Americans	**0.08997**	**0.09366**	0.01558	0.03432	**0.08803**	*	<0.00001	<0.00001
Europeans	**0.04873**	**0.03807**	**0.15448**	**0.08716**	**0.03371**	**0.19028**	*	0.00257
Han Chinese	**0.10812**	**0.0893**	**0.23763**	**0.16215**	0.09715	**0.29154**	0.0677	*

Table 6.13: The results of pairwise exact tests of population differentiation, based on genotype frequencies at the *CYP3A5* locus. The values shown in the Table reflect *p*-values evaluating whether populations differ from each other based on their genotype frequencies. Statistically significant *p*-values following Bonferroni correction are highlighted in pink [adjusted *p*-value = 0.00625; correction for 8 tests].

	Afar	Amhara	Anuak	Maale	Oromo	African-Americans	Europeans	Han Chinese
Afar	*							
Amhara	0.40984	*						
Anuak	**0.00000**	**0.00002**	*					
Maale	**0.00000**	0.06138	**0.00238**	*				
Oromo	0.24611	0.84995	**0.00106**	0.00697	*			
African-Americans	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	*		
Europeans	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00021**	*	
Han Chinese	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00014**	*

Figure 6.12: A Principal Co-ordinates (PCO) plot based on pairwise F_{ST} values between eight populations. Pairwise F_{ST} vales were calculated based on genotype data for an 8063bp *CYP3A5* region. Axis labels show the percentage of genetic distance captured by each axis.



Population codes: AA: African-Americans; AS: Han Chinese; CA: individuals of recent European ancestry; EAF: Afar; EAM: Amhara; EAN: Anuak; EML: Maale; EOR: Oromo.

6.3 Discussion

6.3.1 CYP3A5 variation in Ethiopia is characteristic of sub-Saharan African and non sub-Saharan African populations

A consistent finding of all analyses presented in this thesis is that Ethiopian *CYP3A5* data are characteristic of African and non-African populations. The Ethiopian cohort split into two distinct clusters; with the Afro-Asiatic speaking Afar, Amhara and Oromo clustering closely with non-African populations. Measures of nucleotide diversity for these three Ethiopian groups indicated a skew towards rare variants on a par with the values reported for European and Han Chinese individuals. Principal-components analysis (PCA) found that they cluster closely with Han Chinese individuals. In contrast the remaining populations with recent African ancestry are separate from the Han Chinese and European individuals. There is a well known European contribution to African-American genetic diversity (Reed 1969; Destro-Bisol et al. 1999) and it is interesting that these individuals are furthest away from the individuals of recent European ancestry in the PCO plot (Figure 6.11). Consistent with the previous chapters, population structure across the entire *CYP3A5* gene region in eight populations is explained predominantly by the first principal component and by differences in frequencies of the *CYP3A5*3* allele. Re-sequencing of a larger *CYP3A5* region within diverse Ethiopian populations found that over 30% of all identified variants were singletons; although there were a large number which occurred at high frequency. The allele frequency spectrum in Ethiopians is atypical of what is found for neutral gene regions as there are an excess of variants at either end of the frequency spectrum (see Figure 5.6) as opposed to a steady decline of the number of variants observed at high frequency in the dataset (Jobling et al. 2004).

Gene diversity is lowest in the low/non-expresser haplogroups: *CYP3A5*3* and *CYP3A5*6.* Low measures of gene diversity in Europeans, Han Chinese, the Afar, Amhara and Oromo are a results of high population frequencies of *CYP3A5*3* haplotypes. Ancestral haplotypes are expected to have more diversity than derived. A particularly interesting observation is that the haplotype composition reported here is consistent with that seen in chapters 4; there is a paucity of variation in the gene region on derived haplotypes. Haplotype differentiation is largely defined by variants observed in the gene flanking regions, again demonstrating the paucity of variation observed in the gene sequence. Haplotypes were also defined by high LD between polymorphic sites and suggests that LD across the gene region, and across the *CYP3A* cluster, is high in all eight populations.

Sub-Saharan African populations have been reported to have more diversity than other global populations (Tishkoff and Verrelli 2003) and therefore it is striking that the five Ethiopian populations should be less diverse than African-Americans; and that three of the four Afro-Asiatic speaking groups, are much more characteristic of European and Han Chinese populations, than other Ethiopian populations located ≤900km away. It is possible that Arab migrations into the continent, via the horn of Africa, in the 19th Century (Richards et al. 2003; Kivisild et al. 2004; Lovell et al. 2005) have influenced *CYP3A5* haplotypic structure in the three populations located closest to the Arabian Peninsula (Afar, Amhara and Oromo). Gene flow may also explain the appreciable frequencies of largely African variants *CYP3A5*6* and *CYP3A5*7* in populations from the Arabian Peninsula (reported in chapter 3).

The work presented in this thesis also extends previous work on *CYP3A5* diversity within Ethiopia (Gebeyehu et al.) as it examines *CYP3A5* diversity within Ethiopia according to ethnicity and population sub-structure. The previous study by Gebenyehu *et al* pooled *CYP3A5* data from multiple Ethiopian ethnic groups and so overlooked the substantial inter-ethnic genetic diversity within the region which, as the data presented in this chapter show, is also reflected at the *CYP3A5* locus (Henze 2000; Lovell et al. 2005). Considerable inter-ethnic variability observed in Ethiopia confirms that the criterion for grouping, and analysing individuals, of recent African ancestry by ethnic group and language family in this thesis was correct. The data show that an appreciation of individual Ethiopian population histories will help in the identification of medically relevant genetic variation and identify groups which are likely to be separate from a wider patient cohort.

6.3.2 The potential implications of CYP3A5 variability on clinical outcomes in Ethiopians

The potential medical implications of intra-African diversity in the *CYP3A5* gene were discussed in chapters 3 and 4. The results presented in this chapter focused on intra-Ethiopian diversity and suggest that there are likely to be differences in disease susceptibility and treatment outcomes in diverse groups from the country.

Variability in CYP3A5 expression has been previously shown to affect disease risk (Givens et al. 2003; Dandara et al. 2005; Zhenhua et al. 2005; Bochud et al. 2009) and adverse clinical outcomes associated with drug therapies (Goto et al. 2004). The most significant finding from the work presented in this chapter is that additional variants, which are likely to affect *CYP3A5* transcription and protein expression (see chapter 5), occur almost exclusively on *CYP3A5*1* haplotype backgrounds. A lot of these variants occur at low frequencies but their potential effect on protein expression and consequent medical phenotypes reiterate the

importance of characterising rare, and novel, variation in medically important genes. This is particularly necessary in populations with recent African ancestry (see section 4.5.3) who have been largely underrepresented in medical research, yet are some of the most vulnerable populations to multiple diseases. The haplotype background of these variants suggests that the presence of a *CYP3A5*1* genotype, alone, may not predict CYP3A5 expression profiles.

The results for the entire *CYP3A5* gene also reiterate the importance of not combining Ethiopian populations as a single uniform population; as it is overlooks inter-ethnic diversity, and may exacerbate disease and/or adverse clinical outcomes in specific groups. Furthermore the Afar, Amhara and Oromo are not overly characteristic of sub-Saharan African populations. Throughout this thesis it has been shown that these three populations are characteristic of non sub-Saharan African groups: namely Europeans and Han Chinese individuals. Therefore these three groups cannot be regarded as being typically "African" from a medical perspective and consequently it may not be appropriate to administer drugs, to individuals within these populations, at African-specific dosages. From a medical perspective, these three Ethiopian groups should be considered separately when administering drugs or when calculating disease risk. This will almost certainly change our perspective of population-stratified medicine; by altering our classification of populations as African and non-African, as our understanding of medically important genes, in some of the world's most vulnerable populations, increases.

6.3.3 There are signatures of directional selection at the CYP3A5 locus in Ethiopia

It has previously been reported that there is strong evidence of directional selection, on the *CYP3A5*3* allele, in non sub-Saharan African populations (Thompson et al. 2004; Thompson et al. 2006; Chen et al. 2009). In populations outside of sub-Saharan Africa the *CYP3A5* locus is characterised by an excess of rare variants, high frequencies of the *CYP3A5*3* allele and a paucity of variation on low/non-expresser haplotype backgrounds. An interesting observation is that these characteristics are also true of the Ethiopian *CYP3A5* re-sequencing data.

Tests for departures from neutrality relied on detecting an excess of rare variants. Tajima's *D*, and Fu and Li's *F** and *D** and Fu's *FS*, all indicated a skew towards rare variants in the Ethiopian cohort; although only Fu's *FS* detected a significant departure from neutrality. The power of the Tajima's *D* has been shown to be weak when the number of segregating sites is small (Simonsen et al. 1995) and this may explain why a significant departure from neutrality was not observed in the Ethiopian data. Additionally, Fu and Li's *F** and *D** assume

that a single allele has undergone a selective sweep; this is almost certainly true of populations outside of sub-Saharan Africa. However within Ethiopia, given the high frequencies of *CYP3A5*6* and *CYP3A5*7*, it is possible that there are alternative or additional targets of selection within these populations. As previously reported, there was no evidence of a selective sweep, in Ethiopian populations, based on Fay and Wu's *H* test: although this may be due to the low frequency of derived variants identified from Ethiopian re-sequencing data.

The Ethiopian data are consistent with a hypothesis of positive selection for the *CYP3A5*3* allele in the Afar, Amhara and Oromo. An excess of rare variants is often observed when an allele has been selected for and risen rapidly to high frequency (Sabeti et al. 2006). Signatures of selection are often seen around fixed advantageous alleles; notably a reduction in variation around the selected allele and the creation of an extended region of haplotype homogeneity (Sabeti et al. 2002). The most common low/non-expresser haplotypes from the *CYP3A5*3* and *CYP3A5*6* classes, and the single *CYP3A5*7* haplotype, observed in Ethiopia are all characterised by extended regions of haplotype homogeneity and low levels of variation. This suggests that the *CYP3A5*3* and *CYP3A5*6* alleles may have risen to high frequency rapidly in parts of Ethiopia, perhaps as part of a selective sweep. However, a sweep was not detected from Fay and Wu's *H* test. It is also feasible that *CYP3A5*3* and *CYP3A5*6* were initially neutral variants which then became advantageous in certain Ethiopian groups and so rose rapidly to high frequencies. This phenomenon refers to when a selective sweep occurs on standing neutral variation. Neutral traits will often fluctuate in frequency unlike those which are selected for. Once an allele becomes advantageous, it may rise in frequency and there may be more variation on multiple haplotype backgrounds carrying this variant than seen on non-neutral, highly positive variants.

However it is important to note that analysis of linkage disequilibrium (LD) across the 12,237bp region found that the region extends beyond the *CYP3A5* gene. Therefore it is possible that the *CYP3A5*3* and *CYP3A5*6* alleles are in high LD with an additional or alternative target of selection further up or downstream of the *CYP3A5* gene region itself. Strongly selected alleles will often rise in frequency rapidly along with all tightly linked genetic variation. This often leaves "signatures" of a selective sweep; with high levels of LD across large genomic regions and a paucity of variation (Sabeti et al. 2002). An additional or alternative selection target, which is in high LD with *CYP3A5*3* and/or *CYP3A5*6*, and has risen to high frequency could explain high frequencies of either the *CYP3A5*3* and/or *CYP3A5*6* alleles in Ethiopia. "Allelic surfing" refers to a phenomenon whereby neutral variation tightly linked to a genomic region under selection also rises to high frequency as a long extended haplotype (Hofer et al. 2009).

The following chapter aims to elucidate the evolutionary mechanisms which have shaped diversity in the *CYP3A5* gene; by examining evolutionary relationships between all haplotypes inferred from the Ethiopian and wider African cohorts; estimating the age of low/non-expresser *CYP3A5* alleles to establish whether they are recent mutations; and evaluating the evidence for positive selection on low/non-expresser *CYP3A5* alleles in non-African populations.

7. The recent evolutionary history of *CYP3A5*

7.1 Overview of chapter

7.1.1 Examining the evolutionary relationships between CYP3A5 haplotypes and haplogroups

All inferred *CYP3A5* haplotypes broadly fall into five core haplogroups; each defined by *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles alone or by both *CYP3A5*3* and *CYP3A5*6*. Diversity levels differ between haplogroups which could be the result of different, or multiple, events such as mutation, recombination or selection. *CYP3A5* low/non-expresser haplotypes have a paucity of variation, other than the low/non-expression defining alleles. The mechanism(s) by which haplogroup diversity arose are unclear from the *CYP3A5* haplotype data alone. Haplotype information can be used to infer evolutionary relationships; particularly between individuals within a species (Salzburger et al. 2011).

Differences in population frequencies of *CYP3A5*3*, coupled with extended haplotype homogeneity, has been cited as evidence for positive selection for the low/non-expresser allele in populations furthest from the equator (Thompson et al. 2004). Whilst there are signatures of positive selection for low/non-expresser alleles in populations outside of Africa, the data presented in this thesis suggest that this may also be true for at least three of the five Ethiopian populations; Afar, Amhara and Oromo. However, it is important to differentiate true targets of positive selection, from new, low/non-expresser alleles which have arisen on haplotype backgrounds with little/no variation, and so mimic selection signatures (Nielsen et al. 2009).

The age of an allele is the time since it was created by mutation (Slatkin and Rannala 2000; Rannala and Bertorelle 2001). There are multiple methods by which the age of a specific allele can be estimated. These can be broadly split into phylogenetic and population genetics approaches (Rannala and Bertorelle 2001). A phylogenetic approach aims to estimate the age of an allele by coalescing the time to its most recent common ancestor (MRCA). A representation of inferred evolutionary relationships (in the form of a tree or a network) between each data-point, and the gradual differentiation of all identified chromosomes from an MRCA is created, and the age of the allele is then estimated. A rooted haplotype network, such as one which shows differentiation of the human *CYP3A5* lineage from closely related species such as the chimpanzee (*Pan troglodytes)*, Orang-utan (*Pongo pygmaeus)* and the more distant rhesus macaque (*Macaca mulatta*) can provide a relative chronology; changes close to an ancestral taxon occurred prior to those furthest away (Jobling

et al. 2004). Haplotypes are useful as they incorporate information on markers linked to the allele of interest. However tree-like phylogenies are often unable to account for recombination in lineages; unlike networks which are representations of all possible trees inferred from a given dataset (Jobling et al. 2004).

Microsatellite data are also very useful in estimating allele age; due to their high mutation rates (Wilson and Balding 1998). The stepwise mutation model, frequently used to model microsatellite evolution, assumes that microsatellites mutate at a fixed rate, independent of repeat length, and with the same probability of the number of repeats increasing and decreasing (Kimura and Ohta 1978). However the processes by which microsatellites mutate are often complex and the methods which enable the accurate incorporation of such data into phylogenies; such as stepwise mutation models, are often over simplified (Rannala and Bertorelle 2001). Microsatellite mutation rates differ depending on repeat length and where they are located in the genome (Whittaker et al. 2003; Xu et al. 2005). In the absence of data on the strength of selection at particular loci or the likelihood of mutations occurring at particular microsatellite loci, models such as the stepwise mutation model are useful. The high mutation rate of microsatellites can help to coalesce dates to common ancestors within particular clades and in estimating the mutation rate(s) on particular chromosomes (Wilson and Balding 1998).

Population genetics based approaches incorporate information on recombination and variation linked to a specific mutation of interest to estimate a value of the number of generations, prior to the present, at which a specific mutation is likely to have arisen on a particular haplotype. Initially the underlying population demography is modelled i.e. all possible evolutionary relationships are inferred from the dataset and are modelled in the form of phylogenetic trees/networks. The allele age is inferred from the decay of linkage disequilibrium with nearby markers; either by recombination or mutation, and from the distribution of mutations among descendants from a common ancestor (Rannala and Bertorelle 2001). Both phylogenetics and population-genetics methods aim to establish the underlying causative factors of observed diversity; such as mutation and recombination.

7.1.2 Examining evidence of positive selection at the CYP3A5 locus

In addition to examining how *CYP3A5* haplotype diversity arose in the eight global populations, this chapter aims to evaluate the evidence for positive selection on the *CYP3A5*3* allele in populations from the HapMap phase II and HGDP-CEPH datasets using data extracted from the Haplotter website (http://haplotter.uchicago.edu/) (Voight et al. 2006).

The majority of tests designed to detect positive selection measure the degree of decay of linkage disequilibrium (LD) within the genome using one of two basic statistics; Extended Haplotype Homozygosity (EHH) and Fraction of Recombinant Chromosomes (FRC) (Wang et al. 2006; Huff et al. 2010). Over time LD breaks down over large genomic distances as a result of mutation and recombination. The retention of LD over large distances is a signature of recent positive selection; which results after a selective sweep has driven a selected allele (and all tightly linked variation) to high frequency (Sabeti et al. 2002; Sabeti et al. 2006; Huff et al. 2010). An examination of the decay of LD over large distances in the genome has been applied to whole genome scans to identify large genomic regions which may have been the target of positive selection; although some tests also have sufficient power to detect signatures of selection at single gene loci (Huff et al. 2010).

The LRH test, and EHH statistic, was first described in 2002 by Sabeti (Sabeti et al. 2002). EHH is defined as the probability that two randomly chosen chromosomes, carrying a core haplotype of interest, are identical by descent; i.e. they have the same SNP based haplotype over a pre-defined region. EHH detects the transmission of an extended haplotype that has not undergone recombination (Sabeti et al. 2002; Huff et al. 2010). The LRH test examines the relationship between allele frequencies and the extent of linkage disequilibrium (LD); a selected allele will rise in frequency at a rate that is quicker than recombination is able to break down (Sabeti et al. 2002; Biswas and Akey 2006). The result is a long haplotype characterised by high LD between all markers relative to neutral expectations (Sabeti et al. 2002). The LRH test begins by defining a "core" haplotype; often one associated with a particular SNP. The test then calculates EHH; i.e. the probability that two chromosomes carry the same "core" SNP or haplotype. Positive selection is inferred by identifying core haplotypes which have elevated EHH relative to other haplotypes at the same locus.

An alternative to EHH is the Cross Population Extended Haplotype Homozygosity (XP-EHH) statistic (Sabeti et al. 2007), which has been calculated for populations from the Human Genome Diversity Panel (HGDP). XP-EHH is designed to detect sweeps in which the selected allele has reached fixation in one or more population but remains polymorphic in the global human population (Sabeti et al. 2007). Whereas LRH and iHS detect partial selective sweeps, XP-EHH detects selected alleles that have risen to, or near, fixation in one but not all global populations and as a result is more powerful than LRH or iHS alone (Sabeti et al. 2007).

The iHS statistic was proposed by (Voight et al. 2006) and is also based on the EHH statistic. The iHS test applies to individual SNPs and the test initially calculates a statistic called the integrated EHH (iEHH). If EHH estimates against distance from a core allele are plotted graphically then iEHH is the area under the curve (Voight et al. 2006). The estimates

of iEHH are then standardized by log transformation so that iEHH values have a mean of 0 and a variance of 1. Large negative iHS values (-2.5) indicate unusually long haplotypes carrying the derived allele; whereas large positive values indicate long haplotypes carrying the ancestral allele (Voight et al. 2006). iHS is a standardized statistic which provides a measure of how unusual haplotypes around a core SNP are, relative to the genome as a whole, as is not itself a formal significance test.

A previous study which examined the relative power of multiple statistics used to infer positive selection found that iHS is more powerful in a single-gene candidate test than LRH (Huff et al. 2010). As this thesis is concerned with selection on *CYP3A5,* only iHS data for HapMap phase II and HGDP-CEPH populations are presented in this chapter.

7.2 Specific chapter aims

There are three specific aims of this chapter; the primary aim is to examine the evolutionary relationships between *CYP3A5* haplotypes to infer the mechanism(s) by which low/non-expresser alleles arose on separate chromosomes, from an MRCA. This will be achieved by constructing haplotype networks for each population and haplogroup; an additional aim is to use microsatellite data, for five ethnically diverse Ethiopian populations, to estimate the ages of the most frequent low/non-expresser *CYP3A5* alleles under a stepwise mutation model; the final aim is to examine whether there is evidence of positive selection on the *CYP3A5* gene within populations from HapMap phase II, the Human Genome Diversity Panel (HGDP) and Ethiopia (re-sequenced for this thesis), by differentiating between mimicked signatures of selection (by young and recent mutations) and true signatures.

7.3 Methods

7.3.1 Haplotype networks

Haplotype networks were constructed using Network, version 4.6.1.0 (freely available from http://www.fluxus-engineering.com/netwinfo.htm). The software can construct networks using two different algorithms; median-joining (Bandelt et al. 1999) and reduced median joining (Bandelt et al. 1995). The latter is useful for binary data but median-joining networks are useful for all types of data, for constructing rooted phylogenies/networks and in reducing reticulations. Both algorithms produce networks which represent all shortest possible trees; however this can sometimes produce unnecessary cross links (reticulations). The median-joining algorithm is useful for reducing the number of unnecessary links in a network and was used to infer evolutionary relationships for this thesis. Data on haplotype composition and frequency information were input into the programme and all networks were manually rooted with an ancestral haplotype. The ancestral haplotype was inferred using information, extracted from NCBI (http://www.ncbi.nlm.nih.gov/) on the chimpanzee allele at each polymorphic site. The ancestral haplotype was identical to a *CYP3A5*1* haplotype observed in humans (labelled ROOT in the networks; see section 7.4.1). Networks were created using Network software and then exported to Adobe Photoshop CS4 software for re-colouring.

7.3.2 Estimating the age of common low/non-expresser CYP3A5 alleles

The ages of the *CYP3A5*3, CYP3A5*6* and rs15524 mutations were estimated using microsatellite data obtained for Ethiopians. Microsatellites are highly informative genetic markers due to their high mutation rates which vary depending on the length and size of the repeat (Payseur et al. 2002; Whittaker et al. 2003; Xu et al. 2005) and where they are in the genome. A study examining differences in microsatellite repeat numbers between parents and their offspring (Whittaker et al. 2003) found that the average mutation rate of dinucleotide microsatellite repeats, across the entire genome, is $\sim 4.5 \times 10^{-4}$. This mutation rate was used in the analyses to date *CYP3A5* alleles and is consistent with independent estimates (Farrall and Weeks 1998). However, a mutation rate of $\sim 4.5 \times 10^{-4}$ does not necessarily accurately define the microsatellite mutation rate for the region of the genome which has been genotyped for this chapter. In the absence of the known mutation rate for this region, the use of an average dinucleotide mutation rate may provide an indication of the approximate age of the *CYP3A5* variants genotyped for this thesis.

The gametic phase of *CYP3A5* mutations and the –GT microsatellite (rs72492208) was not determined empirically; therefore the age of each *CYP3A5* mutation was estimated using data for homozygotes for particular variants. As no Ethiopian individual was identified to be a *CYP3A5*7* homozygote, this variant could not be dated. The distribution of microsatellite repeats associated with each of the *CYP3A5*1, CYP3A5*3* and *CYP3A5*6* alleles was plotted. Allele ages were estimated using data for individuals homozygous for particular haplotypes. Under the stepwise mutation model the variance (ASD) in the microsatellite repeat length, from the most recent common ancestor, is a linear function of the mutation rate (μ) and coalescence time in generations (t); ASD=μt (Goldstein et al. 1995; Slatkin 1995).

ASD and t were calculated using Ytime software (Behar et al. 2003). Ytime is a set of Matlab/Octave functions, written by Dr Mike Weale and available at http://www.ucl.ac.uk/tcga/software/. Ytime uses the following formula to estimate ASD:

$$ASD = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}(L_{ij} - L_i^0)^2\right)$$

where the term within the large brackets defines, for a given microsatellite locus *i*, the average difference (over a sample of *n* chromosomes) in the repeat size between each sampled chromosome (L_{ij}) and the repeat size of the MRCA (L_i^0). The complete formula, therefore, gives the average of this average over all microsatellite loci sampled.

Ytime estimates the Time to the Most Recent Common Ancestor (TMRCA) in a clade for which microsatellite data are available. The microsatellite length of the ancestral MRCA is assumed to be known. For this study the length of the microsatellite on the ancestral haplotype was assumed to be 35; as the majority of *CYP3A5*1* haplotypes had 35 repeats. Confidence intervals for the age estimates were obtained from calculating the distances between the ancestral and derived chromosomes under a star-genealogy model; based on the results of network analysis of *CYP3A5* haplotypes (see section 7.4.1). A generation was assumed to be 32 years, based on the average estimate for females (~29 years) and males (~35 years) (Tremblay and Vezina 2000).

7.3.3 Testing for selection at the CYP3A5 locus

There are a number of different methods which aim to identify genomic regions which are likely to have been targets of positive selection (see section 7.1.2). In addition to those

used to examine departures from neutrality in chapter 4 and 6, this thesis utilises iHS data for the *CYP3A5* gene available online for HapMap phase II and HGDP-CEPH populations.

Data extracted from Haplotter provided a visual comparison of iHS estimates for genomic regions surrounding the *CYP3A5* gene and around the *CYP3A5*3* allele in different populations. The data were then compared with those for other genes for which positive selection is known to have acted.

iHS estimates could not be accurately obtained for the Ethiopian data due to the smaller size of the re-sequenced region (~13kb) and the method of re-sequencing meant that large intronic regions, that are part of *CYP3A5,* could not be assessed and this may influence the iHS estimates.

7.4 Results

7.4.1 Network analysis of CYP3A5 whole gene haplotypes

Networks were constructed for *CYP3A5* haplotypes inferred for an 8063bp region in 8 populations (see Figure 7.1 and the corresponding key in Table 7.1). *CYP3A5* haplotypes are the taxa and are connected by branches. The size of each taxon corresponds to its global frequency. Taxa are coloured coded according to which haplogroup they belong. *CYP3A5*1* haplotypes are coloured in yellow; *CYP3A5*3* in red, *CYP3A5*6* in blue; *CYP3A5*7* in green and *CYP3A5*3/*6* recombinants in purple. The haplotype named "ROOT" was used to manually root the network and corresponds to the *CYP3A5* haplotype with no mutations identified across the gene region (Table 7.1). This haplotype is also identical to that observed in the closely related chimpanzee sequence.

A star like network phylogeny was inferred from the *CYP3A5* haplotype data. Star phylogenies are consistent with rapid growth and differentiation of haplotypes from an MRCA, or with a selective sweep. Many expresser haplotypes were rare in the global dataset and the modal haplotypes were low/non-expresser; each 1-2 point mutations away from the MRCA. The expresser haplotype is the ancestral state and high frequencies of low/non-expresser haplotypes, coupled with an inferred star phylogeny, is consistent with a hypothesis that there has been a rapid increase in *CYP3A5*3* and *CYP3A5*6* allele and haplotype frequencies. The network confirms the results of gene diversity analyses (see chapter 5); diversity is higher in the *CYP3A5*1* haplogroup than in the low/non-expresser haplogroups. Of the low/non-expresser haplogroups; diversity is highest in the *CYP3A5*3* cluster.

Figure 7.1: A haplotype network of all inferred *CYP3A5* haplotypes from an 8063bp region in 8 populations (Ethiopians and Coriell populations). Networks assume single mutational steps. Haplotypes are coloured according to the haplogroup to which they belong; *CYP3A5*3* haplotypes are shown in red; *CYP3A5*1* in yellow; *CYP3A5*7* in green; *CYP3A5*6* in blue and *CYP3A5*3/*6* recombinant haplotypes are shown in purple. The size of each haplotype is proportional to its frequency in the global database. Table 7.1 contains information on the exact composition of each coded haplotype in accordance with Figure 7.1.

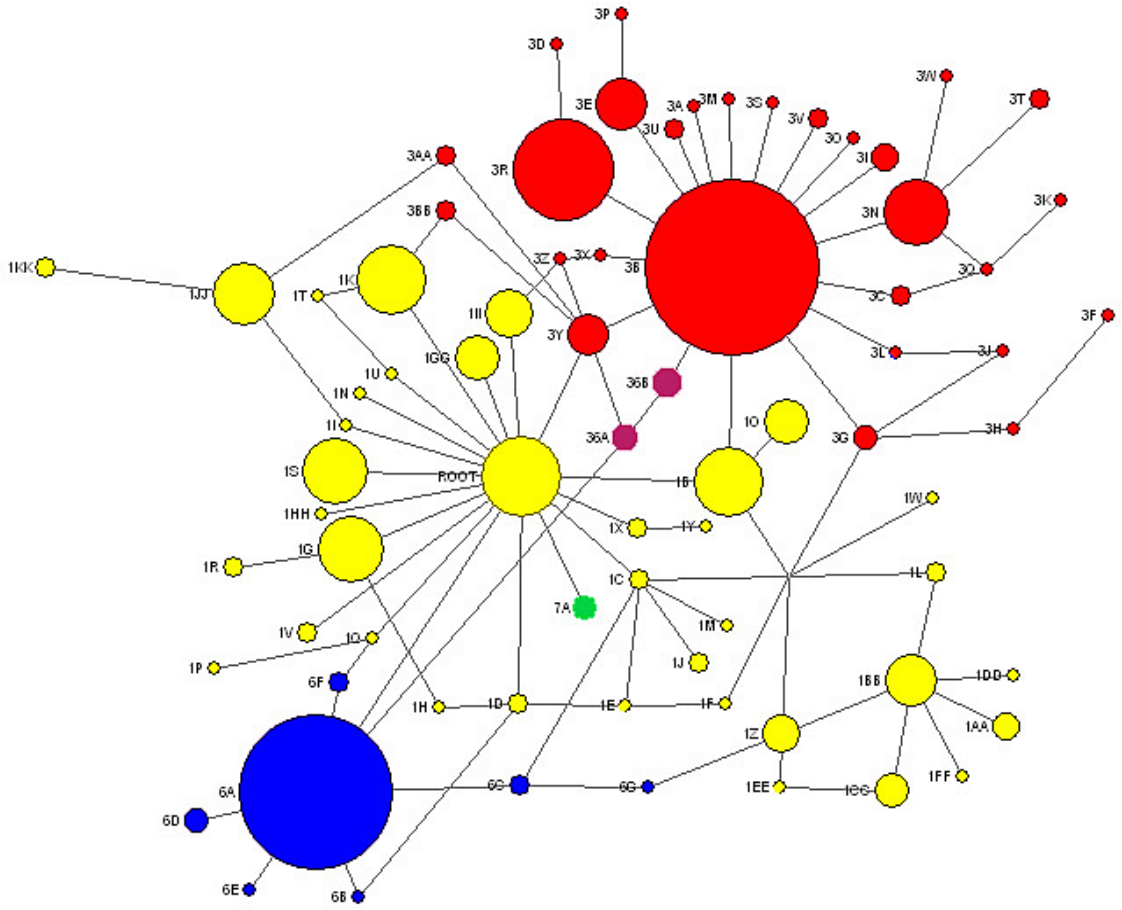
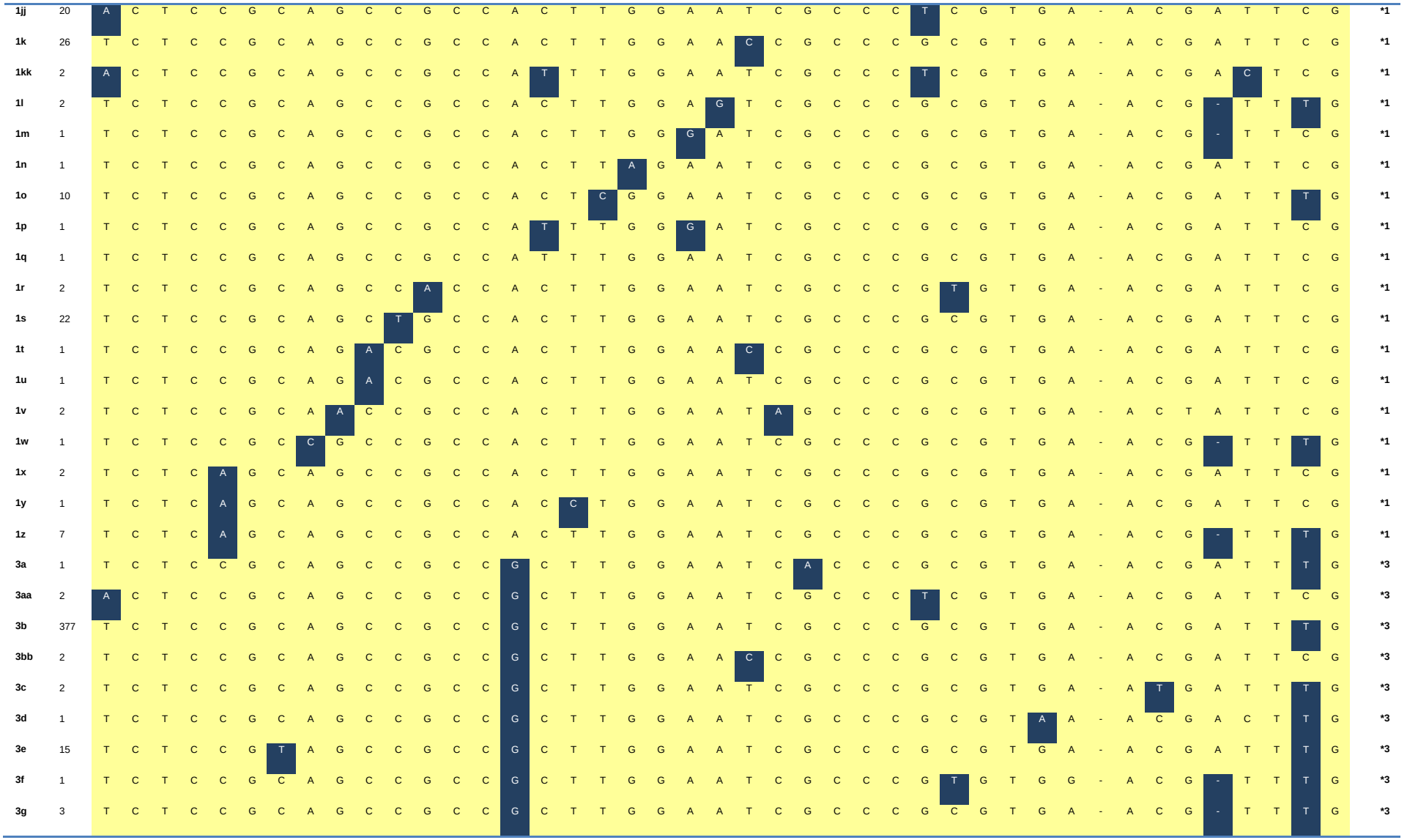
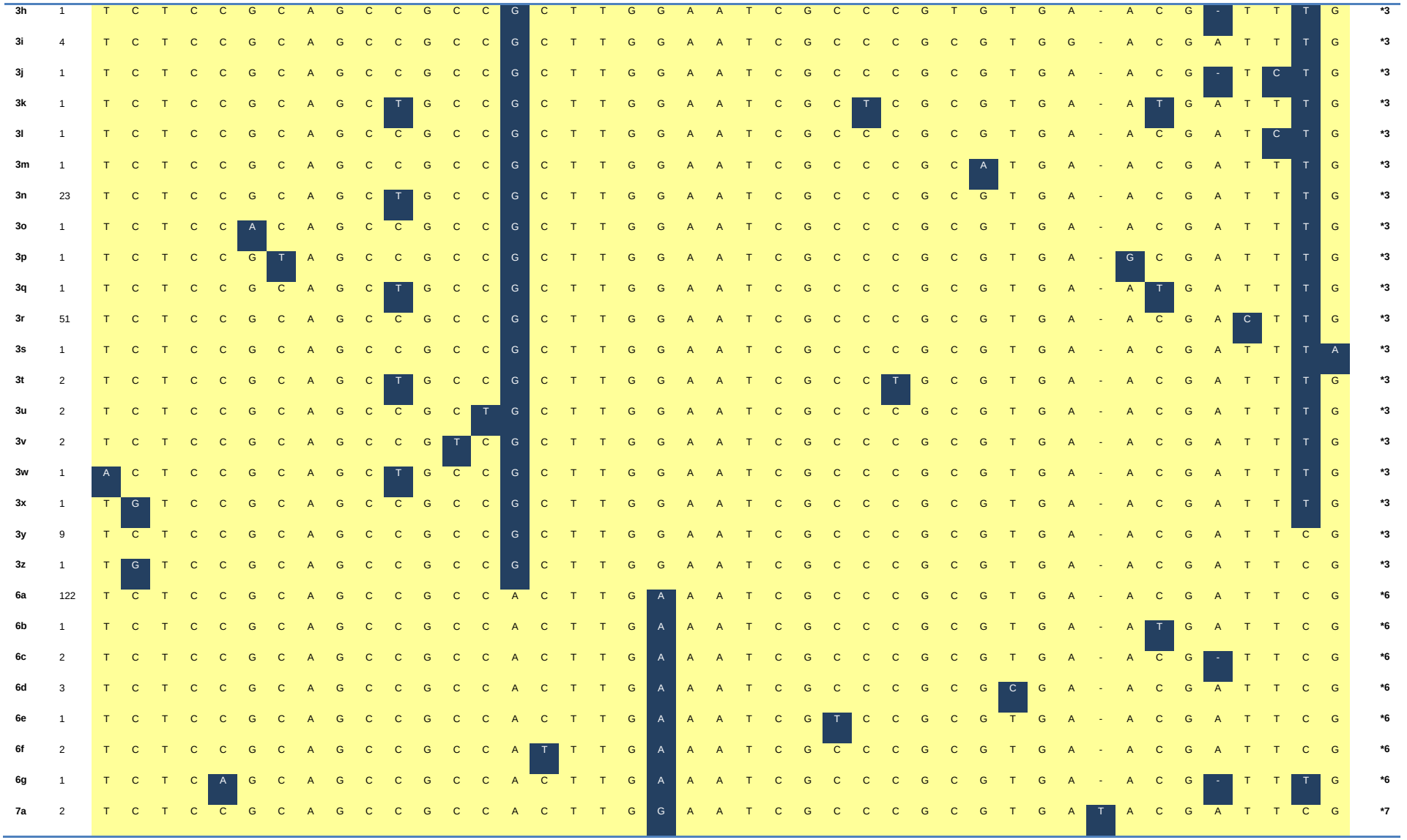


Table 7.1: A table showing the composition of haplotypes analysed in Figure 7.1. The positions of polymorphic sites (numbered from the ATG start codon where base A is +1) are shown along the top. The codes correspond to those used in Figure 7.1 and *n* is the frequency of each haplotype in the global dataset. Haplogroup refers to whether each haplotype belongs to the *CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*7* or *CYP3A5*3/*6* cluster of haplotypes. Ancestral alleles at each nucleotide positions are coloured in yellow and derived alleles in blue.

Code	*n*	-795	-748	-709	-705	-551	-86	-74	-15	127	182	5209	5229	5244	5416	6980	7201	7354	13167	13370	14684	14714	14877	15782	15936	16785	17017	17020	17112	17157	18995	19159	19203	26943	27044	27128	27201	27520	29872	30016	31493	31545	31605	31812	Haplogroup
36a	3	T	C	T	C	C	G	C	A	G	C	C	G	C	C	G	C	T	T	G	A	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	*3/*6
36b	3	T	C	T	C	C	G	C	A	G	C	C	G	C	C	G	C	T	T	G	A	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	T	G	*3/*6
1a	32	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	*1
1aa	4	T	C	T	C	A	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	G	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	T	G	*1
1b	25	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	T	G	*1
1bb	15	T	C	T	C	A	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	G	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	T	G	*1
1c	2	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	C	G	*1
1cc	6	T	C	T	C	A	G	C	A	G	C	C	G	C	C	A	T	T	T	G	G	A	G	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	T	G	*1
1d	2	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	T	G	A	T	T	C	G	*1
1dd	1	T	C	T	C	A	G	C	A	A	C	C	G	C	C	A	C	T	T	G	G	A	G	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	T	G	*1
1e	1	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	T	G	-	T	T	C	G	*1
1ee	1	T	C	T	C	A	G	C	A	G	C	C	G	C	C	A	T	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	T	G	*1
1f	1	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	T	G	-	T	T	T	G	*1
1ff	1	T	C	T	C	A	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	G	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	C	G	*1
1g	22	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	T	G	T	G	A	-	A	C	G	A	T	T	C	G	*1
1gg	11	T	C	T	-	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	*1
1h	1	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	T	G	T	G	A	-	A	T	G	A	T	T	C	G	*1
1hh	1	T	C	C	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	*1
1i	1	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	T	C	G	T	G	A	-	A	C	G	A	T	T	C	G	*1
1ii	13	T	G	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	G	C	C	C	G	C	G	T	G	A	-	A	C	G	A	T	T	C	G	*1
1j	2	T	C	T	C	C	G	C	A	G	C	C	G	C	C	A	C	T	T	G	G	A	A	T	C	A	C	C	C	G	C	G	T	G	A	-	A	C	G	-	T	T	C	G	*1


1jj 20 A C T C C G C A G C C G C C A C T T G G A A T C G C C C T C G T G A - A C G A T T C G *1
1k 26 T C T C C G C A G C C G C C A C T T G G A A C C G C C C G C G T G A - A C G A T T C G *1
1kk 2 A C T C C G C A G C C G C C A T T T G G A A T C G C C C T C G T G A - A C G A C T C G *1
1l 2 T C T C C G C A G C C G C C A C T T G G A G T C G C C C G C G T G A - A C G - T T T G *1
1m 1 T C T C C G C A G C C G C C A C T T G G G A T C G C C C G C G T G A - A C G - T T C G *1
1n 1 T C T C C G C A G C C G C C A C T T A G A A T C G C C C G C G T G A - A C G A T T C G *1
1o 10 T C T C C G C A G C C G C C A C T C G G A A T C G C C C G C G T G A - A C G A T T T G *1
1p 1 T C T C C G C A G C C G C C A T T T G G G A T C G C C C G C G T G A - A C G A T T C G *1
1q 1 T C T C C G C A G C C G C C A T T T G G A A T C G C C C G C G T G A - A C G A T T C G *1
1r 2 T C T C C G C A G C C A C C A C T T G G A A T C G C C C G T G T G A - A C G A T T C G *1
1s 22 T C T C C G C A G C T G C C A C T T G G A A T C G C C C G C G T G A - A C G A T T C G *1
1t 1 T C T C C G C A G A C G C C A C T T G G A A C C G C C C G C G T G A - A C G A T T C G *1
1u 1 T C T C C G C A G A C G C C A C T T G G A A T C G C C C G C G T G A - A C G A T T C G *1
1v 2 T C T C C G C A A C C G C C A C T T G G A A T A G C C C G C G T G A - A C T A T T C G *1
1w 1 T C T C C G C C G C C G C C A C T T G G A A T C G C C C G C G T G A - A C G - T T T G *1
1x 2 T C T C A G C A G C C G C C A C T T G G A A T C G C C C G C G T G A - A C G A T T C G *1
1y 1 T C T C A G C A G C C G C C A C C T G G A A T C G C C C G C G T G A - A C G A T T C G *1
1z 7 T C T C A G C A G C C G C C A C T T G G A A T C G C C C G C G T G A - A C G - T T T G *1
3a 1 T C T C C G C A G C C G C C G C T T G G A A T C A C C C G C G T G A - A C G A T T T G *3
3aa 2 A C T C C G C A G C C G C C G C T T G G A A T C G C C C T C G T G A - A C G A T T C G *3
3b 377 T C T C C G C A G C C G C C G C T T G G A A T C G C C C G C G T G A - A C G A T T T G *3
3bb 2 T C T C C G C A G C C G C C G C T T G G A A C C G C C C G C G T G A - A C G A T T C G *3
3c 2 T C T C C G C A G C C G C C G C T T G G A A T C G C C C G C G T G A - A T G A T T T G *3
3d 1 T C T C C G C A G C C G C C G C T T G G A A T C G C C C G C G T A A - A C G A C T T G *3
3e 15 T C T C C G T A G C C G C C G C T T G G A A T C G C C C G C G T G A - A C G A T T T G *3
3f 1 T C T C C G C A G C C G C C G C T T G G A A T C G C C C G T G T G G - A C G - T T T G *3
3g 3 T C T C C G C A G C C G C C G C T T G G A A T C G C C C G C G T G A - A C G - T T T G *3


3h 1 TCTCCGCAGCCGCCGCTTGGAATCGCCCGTGTGA-ACG-TTTG *3
3i 4 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGGG-ACGATTTG *3
3j 1 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACG-TCTG *3
3k 1 TCTCCGCAGCTGCCGCTTGGAATCGCTCGCGTGA-ATGATTTG *3
3l 1 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGATCTG *3
3m 1 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCATGA-ACGATTTG *3
3n 23 TCTCCGCAGCTGCCGCTTGGAATCGCCCGCGTGA-ACGATTTG *3
3o 1 TCTCCACAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGATTTG *3
3p 1 TCTCCGTAGCCGCCGCTTGGAATCGCCCGCGTGA-GCGATTTG *3
3q 1 TCTCCGCAGCTGCCGCTTGGAATCGCCCGCGTGA-ATGATTTG *3
3r 51 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGCTTTG *3
3s 1 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGATTTA *3
3t 2 TCTCCGCAGCTGCCGCTTGGAATCGCCTGCGTGA-ACGATTTG *3
3u 2 TCTCCGCAGCCTGCGCTTGGAATCGCCCGCGTGA-ACGATTTG *3
3v 2 TCTCCGCAGCCGTCGCTTGGAATCGCCCGCGTGA-ACGATTTG *3
3w 1 ACTCCGCAGCTGCCGCTTGGAATCGCCCGCGTGA-ACGATTTG *3
3x 1 TGTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGATTTG *3
3y 9 TCTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGTTCG *3
3z 1 TGTCCGCAGCCGCCGCTTGGAATCGCCCGCGTGA-ACGTTCG *3
6a 122 TCTCCGCAGCCACTTGAAATCGCCCGCGTGA-ACGTTCG *6
6b 1 TCTCCGCAGCCACTTGAAATCGCCCGCGTGA-ATGTTCG *6
6c 2 TCTCCGCAGCCACTTGAAATCGCCCGCGTGA-ACG-TTCG *6
6d 3 TCTCCGCAGCCACTTGAAATCGCCCGCGCGAA-ACGATTCG *6
6e 1 TCTCCGCAGCCACTTGAAATCGTCCGCGTGA-ACGATTCG *6
6f 2 TCTCCGCAGCCACTTAAATCGCCCGCGTGA-ACGATTCG *6
6g 1 TCTCAGCAGCCACTTGAAATCGCCCGCGTGA-ACG-TTTG *6
7a 2 TCTCCGCAGCCACTTGGAATCGCCCGCGTGAACGATTCG *7

There is evidence of recombination which is seen as reticulations in the network. Recombination has occurred mostly between expresser (*CYP3A5*1*) and low/non-expresser haplotypes from the *CYP3A5*3* and *CYP3A5*6* haplogroups; which manifested itself in haplotypes 3AA, 3BB, 3Z, 6B, 6F and 6G. There is also evidence of recombination between the high frequency *CYP3A5*3* and *CYP3A5*6* haplotypes, resulting in haplotypes 36A and 36B. The majority of recombination events have occurred in Ethiopian and African-American populations (see Figure 7.2). *CYP3A5*6* appears to have arisen on three different *CYP3A5*1* haplotypes. The two recombinant *CYP3A5*3/CYP3A5*6* haplotypes are separated by a single point mutation; it appears that *CYP3A5*6* evolved twice on two separate *CYP3A5*3* backgrounds generating two recombinant haplotypes. *CYP3A5*7* is a point mutation away from the root haplotype.

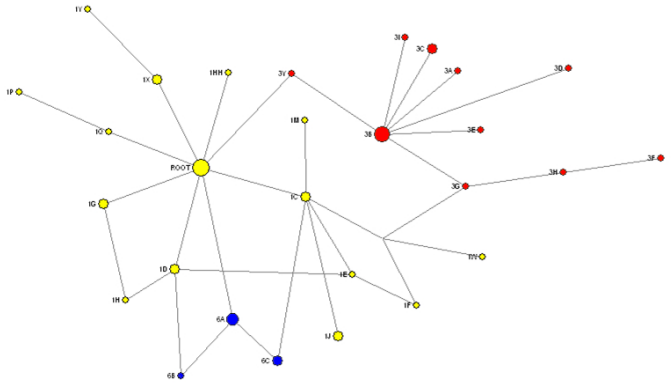
7.4.2. Network analysis of entire gene haplotypes by population

Nearly all haplotypes observed in Europeans and Han Chinese individuals are defined by the *CYP3A5*3* mutation. A number of private *CYP3A5*3* haplotypes were observed in each population (see Figure 7.2); although they were observed at low frequencies. Consistent with the results presented in chapter 6, all populations with recent African ancestry had more differentiation of *CYP3A5* haplogroups; although surprisingly the Anuak had low levels of diversity in the *CYP3A5*3* haplogroup. Recombination was observed in all populations with recent African ancestry and in Han Chinese individuals; suggesting that *CYP3A5* haplotype diversity within these populations is a result of both mutation and recombination events. In contrast, European haplotypes arose exclusively by mutation alone.

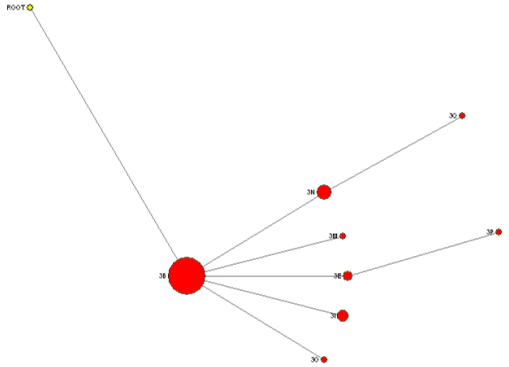
No one individual, from the global cohort, was homozygous for the ancestral "ROOT" haplotype. The modal haplotype in every population was the derived *CYP3A5*3* (3B) haplotype. Within populations with recent African ancestry, a number of rare *CYP3A5*6* haplotypes were only observed in the Anuak and Maale. A comparison of a larger genomic region in all populations may lead to further differentiation of haplotypes. When the Coriell populations were re-sequenced across the *CYP3A* cluster; diversity was lowest in the *CYP3A5* gene; and *CYP3A* haplotypes which are defined by the *CYP3A5*3* mutation are characterised by distinctive LD patterns (Thompson et al. 2006). Little differentiation of *CYP3A5* haplotypes is seen; consistent with the low nucleotide diversity estimates.

Figure 7.2: Networks of inferred *CYP3A5* haplotypes from an 8063 base pair region within 8 populations. Networks assume single mutational steps. Haplotypes are coloured according to the haplogroup to which they belong; *CYP3A5*3* haplotypes are shown in red; *CYP3A5*1* in yellow; *CYP3A5*7* in green; *CYP3A5*6* in blue and *CYP3A5*3/*6* recombinant haplotypes are shown in purple. The size of each haplotype is proportional to its frequency in the global database. Table 7.1 contains information on the exact composition of each coded haplotype in accordance with Figure 7.1. The haplotype named "ROOT" was used to root each network and is equivalent to the chimpanzee *CYP3A5* sequence and to haplotype 1 (Table 7.1).

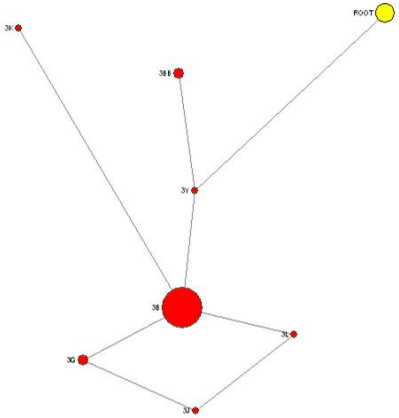
African-Americans



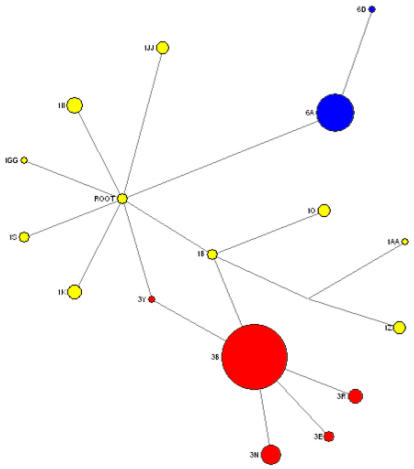
Europeans



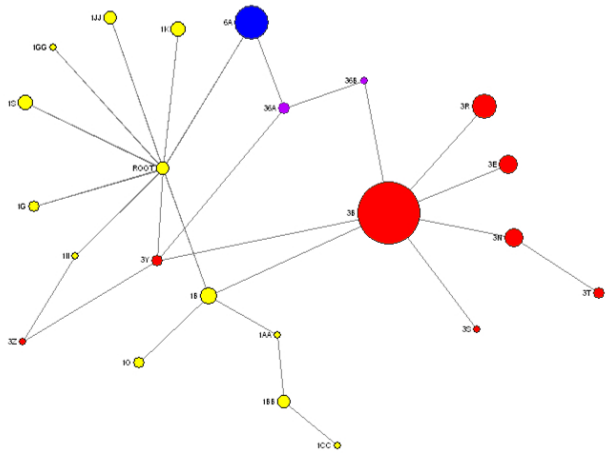
Han Chinese



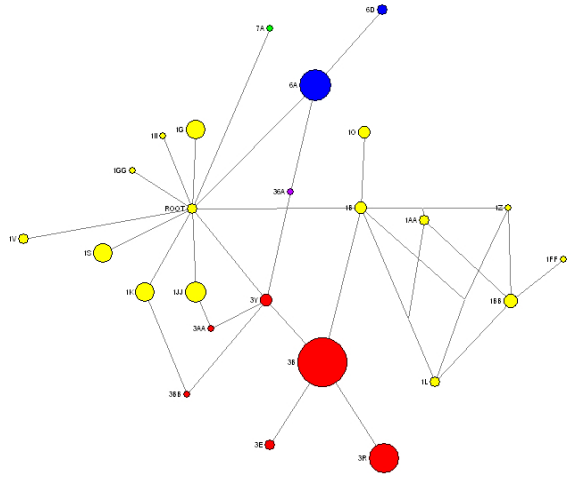
Afar



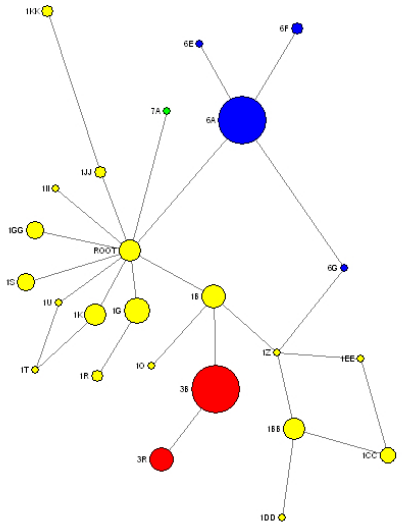
Amhara



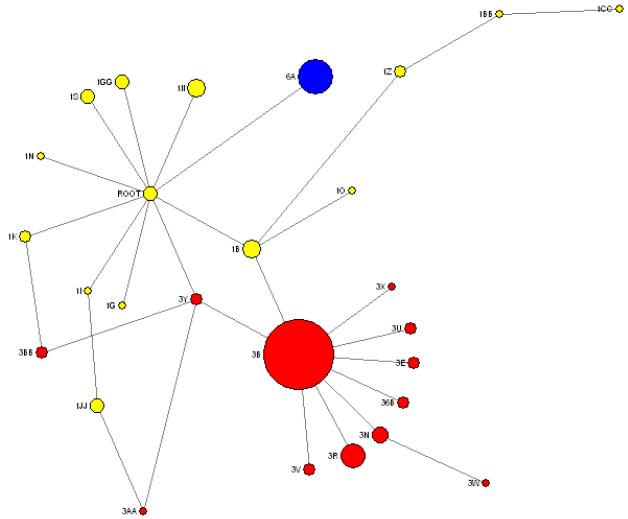
Maale



Anuak



Oromo



7.4.3 Examining the evolutionary relationships between CYP3A5 haplotypes inferred for a ~4kb region in sixteen populations

A network of all haplotypes, inferred for a 4448bp region, in thirteen African populations is presented in Figure 7.3; all haplotypes inferred for a 4006bp region in sixteen populations (see chapter 4) in Figure 7.4; and by geographic region in Figure 7.5. The corresponding key for all haplotypes in each network is presented in Table 7.2.

The modal *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* haplotypes are defined by a paucity of variation. It is possible that there are additional variants, found on *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* haplotype backgrounds, within a larger gene region in these populations. A single, synonymous, exonic variant (observed in sub-Saharan Africa) was inferred on a *CYP3A5*1* haplotype background. The five individuals (four Afro-Asiatic speakers and one Niger-Congo B speaker) identified as having a novel intronic ten base pair deletion (see chapter 4) were all *CYP3A5*3* homozygotes. This suggests that a novel variant, likely to affect *CYP3A5* transcription, has evolved onto an existing low/non expresser haplotype background in West Central Africa (haplotype 18). It is possible that *CYP3A5*3* haplotypes are more diverse, over a larger genomic region, in Niger-Congo populations than in Europeans and Han Chinese analysed for this thesis.

As for a larger *CYP3A5* region, see sections 7.4.1-7.4.2, star-phylogenies were inferred from the haplotype data; consistent with rapid population growth or of a selective sweep (Martins and Housworth 2002; Jobling et al. 2004). Single point mutations separate each haplotype and there was evidence of recombination which manifested itself in haplotypes 3, 9, 18, 21, 24, 25, 27 and 29.

Across geographic regions (Figure 7.5) diversity was highest in East Africa and West Central Africa. Overall African populations were more diverse than populations from the Coriell Repositories (diverse populations sampled from North America). *CYP3A5*3* haplotype diversity is higher in Coriell populations than in Africa (with the exception of East Africa). *CYP3A5*6* haplogroup differentiation is greatest in West Central Africa; whereas *CYP3A5*7* is consistent across regions in which it was observed, perhaps due to this allele having arisen from a recent mutation. West Central African diversity is likely to be influenced by the presence of both Afro-Asiatic and Niger-Congo speaking groups, sampled from the region; groups which differ from each other in terms of *CYP3A5* variation (see chapters 3 and 4). Across Niger-Congo speaking groups are homogeneous in terms of haplotype diversity and structure is similar; which is due to shared recent ancestry as a result of the Bantu expansion (Beleza et al. 2005; Berniell-Lee et al. 2009).

Figure 7.3: A network of all haplotypes inferred for a 4448 base pair *CYP3A5* region, in thirteen African populations. Networks assume single mutational steps. The size of each haplotype (pies) is proportional to its frequency in the African cohort. Haplotype codes refer to those listed in Table 7.2. "Root" refers to the ancestral sequence for the polymorphic sites; inferred from the closely related chimpanzee sequence; and identical to haplotype 1. Haplotypes defined by the *CYP3A5*3* mutation are coloured in red; those defined by *CYP3A5*6* in blue; *CYP3A5*1* in yellow; *CYP3A5*7* in green; and *CYP3A5*3/*6* recombinant in purple.

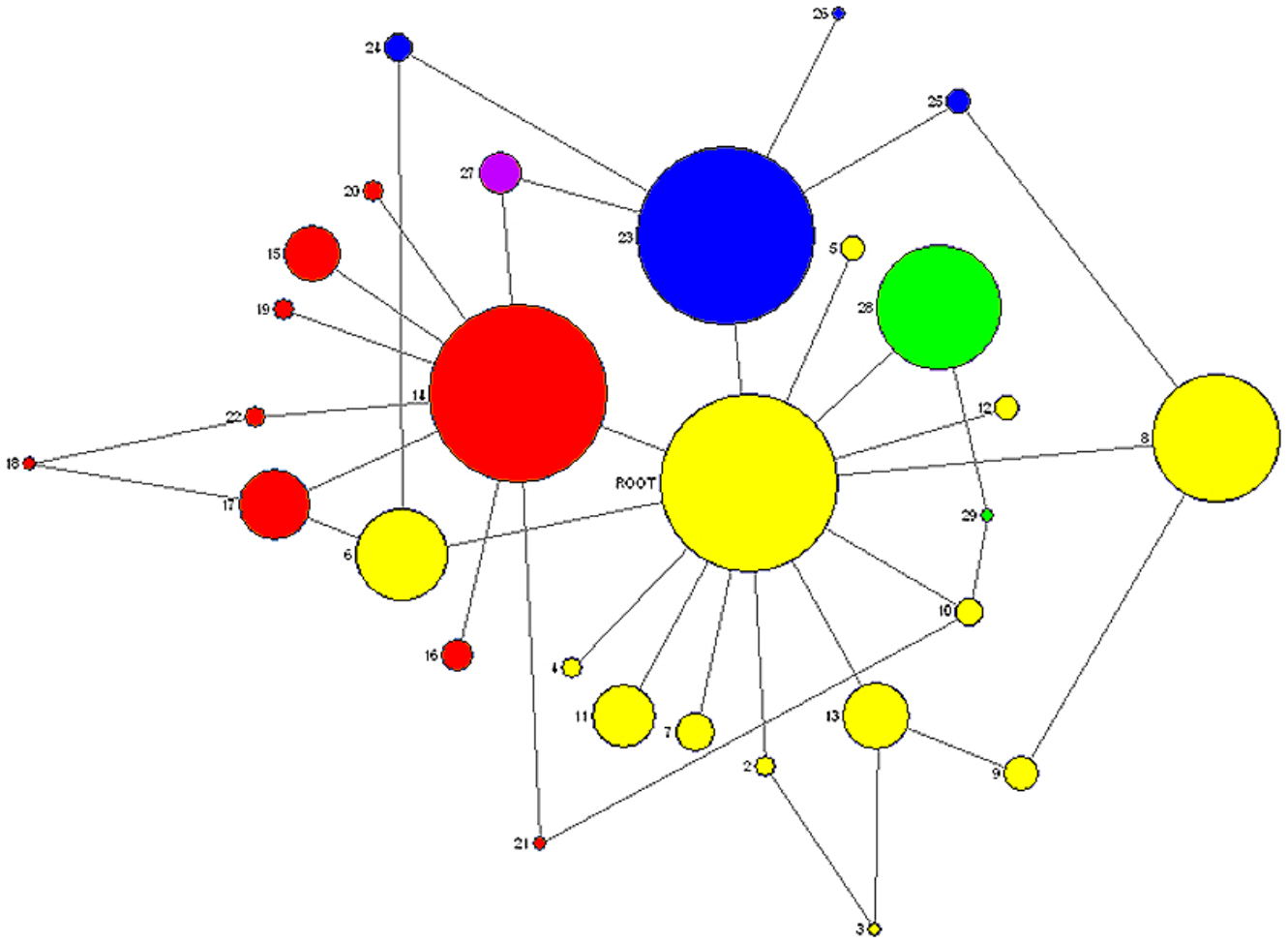


Figure 7.4: A network of all haplotypes inferred for a 4006 base pair *CYP3A5* region in sixteen populations. Networks assume single mutational steps. The size of each haplotypes (pies) is proportional its frequency in the global cohort. Haplotype codes refer to those listed in Table 7.2. "Root" refers to the ancestral sequence for the polymorphic sites; inferred from the closely related chimpanzee sequence; and identical to haplotype 1. Haplotypes defined by the *CYP3A5*3* mutation are coloured in red; those defined by *CYP3A5*6* in blue; *CYP3A5*1* in yellow; *CYP3A5*7* in green; and *CYP3A5*3/*6* recombinant in purple.

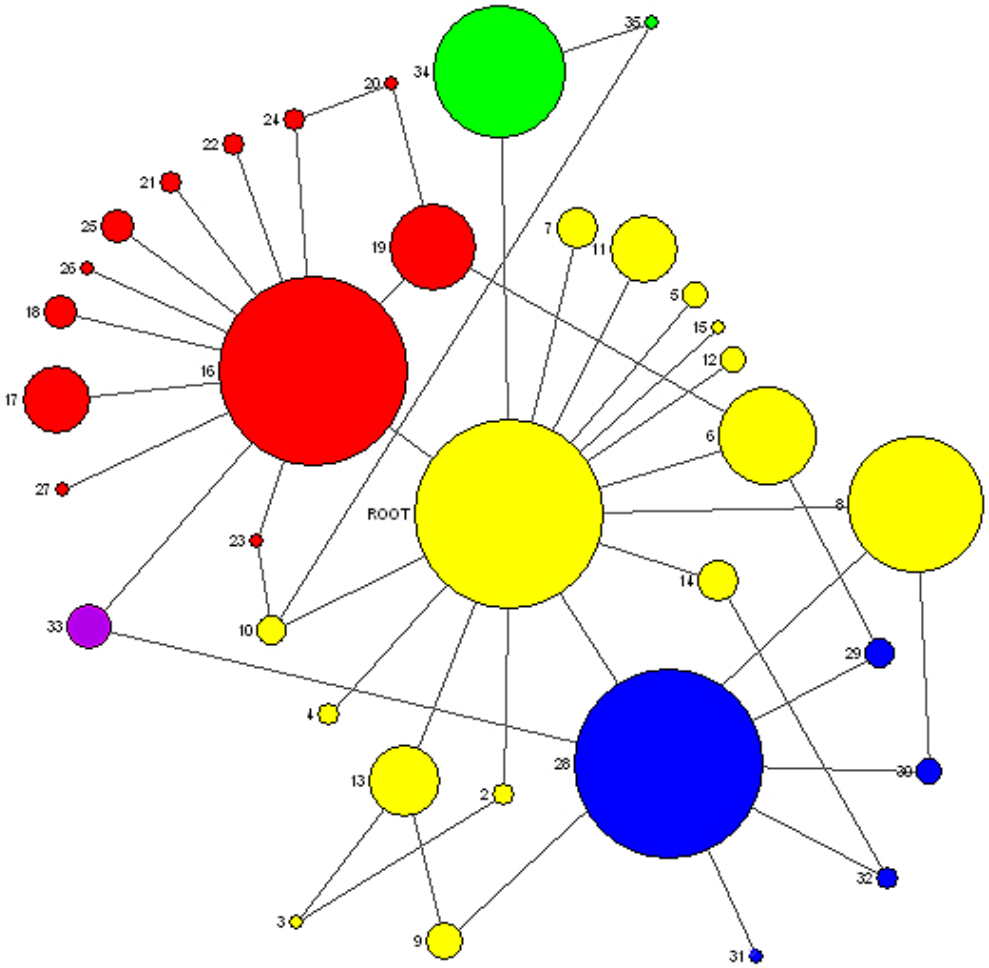
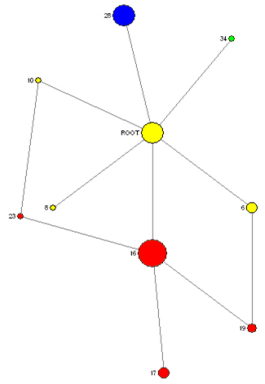
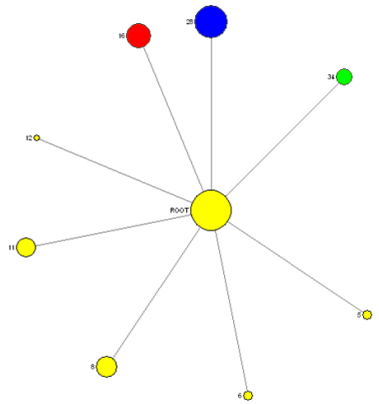


Figure 7.5: Networks of haplotypes within each geographic region. Population-specific haplotype frequencies are presented in Figures 7.1b-g. The size of each haplotypes (pies) is proportional its frequency in the global cohort. Haplotype codes refer to those listed in Table 7.2. "Root" refers to the ancestral sequence for the polymorphic sites; inferred from the closely related chimpanzee sequence; and identical to haplotype 1. Haplotypes defined by the *CYP3A5*3* mutation are coloured in red; those defined by *CYP3A5*6* in blue; *CYP3A5*1* in yellow; *CYP3A5*7* in green; and *CYP3A5*3/*6* recombinant in purple. "*n*" is the number of chromosomes.

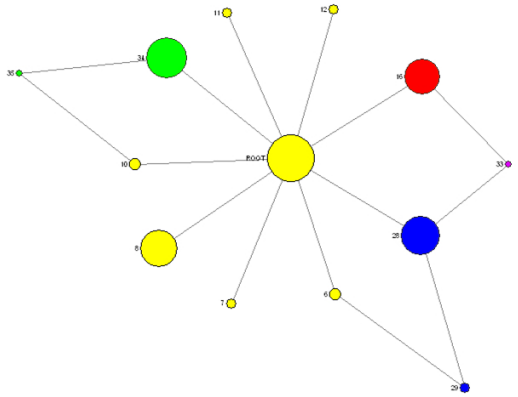
a) North Africa (*n*=56)



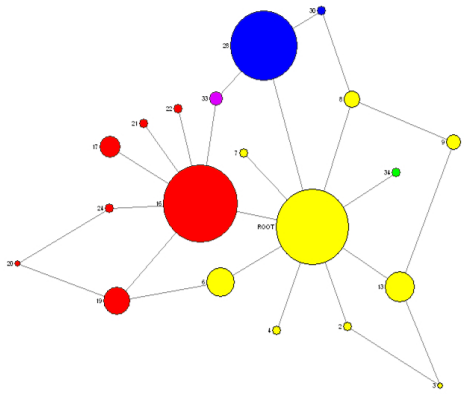
c) West Africa (*n*=107)



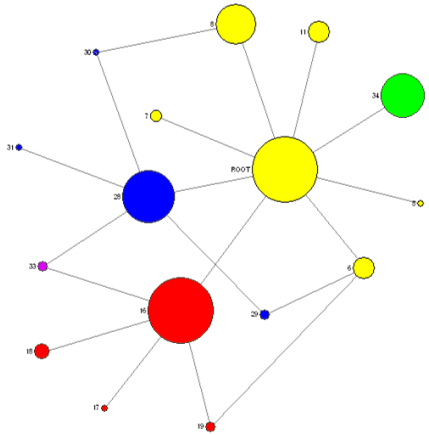
e) South East Africa (*n*=192)



b) East Africa (*n*=761)



d) West Central Africa (*n*=357)



f) North American Coriell samples (*n*=138)

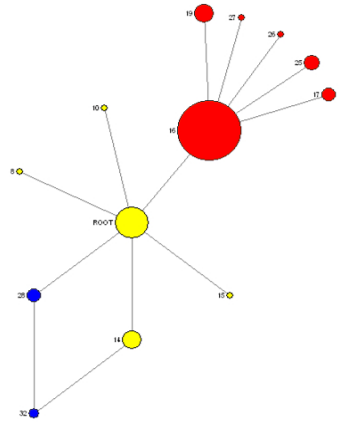


Table 7.2: A table showing the composition of haplotypes analysed in Figures 7.3-7.5. The positions of polymorphic sites (numbered from the ATG start codon where base A is +1) are shown along the top. The codes correspond to those used in Figures 7.3-7.5 and *n* is the frequency of each haplotype in the global dataset. Haplogroup refers to whether each haplotype belongs to the *CYP3A5*1, CYP3A5*3, CYP3A5*6, CYP3A5*7* or *CYP3A5*3/*6* cluster of haplotypes. Ancestral alleles at each nucleotide positions are coloured in yellow and derived alleles in blue.

Haplotype class	Code	-86	-74	-15	74	127	136	182	289	318	5209	5229	5244	5416	6980	7201	7354	7355	14684	14714	14830	14877	26943	27044	27128	N
*1	1	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	400
	2	G	C	A	I	**A**	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	2
	3	G	C	A	I	**A**	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	**G**	G	A	-	1
	4	G	C	A	I	G	C	**A**	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	2
	5	G	C	A	I	G	C	C	**C**	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	3
	6	G	C	A	I	G	C	C	G	G	**T**	G	C	C	A	C	T	C	G	A	C	A	G	A	-	41
	7	G	C	A	I	G	C	C	G	G	C	**A**	C	C	A	C	T	C	G	A	C	A	G	A	-	7
	8	G	C	A	I	G	C	C	G	G	C	G	C	C	A	**T**	T	C	G	A	C	A	G	A	-	86
	9	G	C	A	I	G	C	C	G	G	C	G	C	C	A	**T**	T	C	G	A	C	**G**	G	A	-	6
	10	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	**C**	C	G	A	C	A	G	A	-	5
	11	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	**G**	C	A	G	A	-	21
	12	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	**T**	A	G	A	-	3
	13	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	**G**	G	A	-	23
	14	G	C	A	I	G	C	C	G	**A**	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	7
	15	G	C	**C**	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	-	1
*3	16	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	608
	17	G	**T**	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	20
	18	G	C	A	**D**	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	5
	19	G	C	A	I	G	C	C	G	G	**T**	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	33
	20	G	C	A	I	G	C	C	G	G	**T**	G	C	C	**G**	C	T	**T**	G	A	C	A	G	A	-	1
	21	G	C	A	I	G	C	C	G	G	C	G	**T**	C	**G**	C	T	C	G	A	C	A	G	A	-	2
	22	G	C	A	I	G	C	C	G	G	C	G	C	**T**	**G**	C	T	C	G	A	C	A	G	A	-	2
	23	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	**C**	C	G	A	C	A	G	A	-	1
	24	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	**T**	G	A	C	A	G	A	-	2
	25	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	**G**	-	5
	26	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	**A**	A	-	1
	27	**A**	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	G	A	C	A	G	A	-	1
*6	28	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	256
	29	G	C	A	I	G	C	C	G	G	**T**	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	4
	30	G	C	A	I	G	C	C	G	G	C	G	C	C	A	**T**	T	C	**A**	A	C	A	G	A	-	3
	31	G	C	A	I	G	**T**	C	G	G	C	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	1
	32	G	C	A	I	G	C	C	G	**A**	C	G	C	C	A	C	T	C	**A**	A	C	A	G	A	-	2
*3/*6	33	G	C	A	I	G	C	C	G	G	C	G	C	C	**G**	C	T	C	**A**	A	C	A	G	A	-	9
*7	34	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	T	C	G	A	C	A	G	A	**T**	76
	35	G	C	A	I	G	C	C	G	G	C	G	C	C	A	C	**C**	C	G	A	C	A	G	A	**T**	1

7.4.4. Network analysis of CYP3A5 haplogroups

Network analysis of *CYP3A5* haplogroups is presented in Figures 7.6-7.7. Overall, for both the entire *CYP3A5* gene, and a 4kb region, the greatest differentiation of haplotypes was observed in the *CYP3A5*1* and *CYP3A5*3* haplogroups. Recombination was observed in the *CYP3A5*1* and *CYP3A5*3* haplogroups suggesting that diversity, within both haplogroups, arose by mutation and recombination events. Star phylogenies were inferred for the *CYP3A5*1* and *CYP3A5*3* haplogroups also, and the *CYP3A5*3* network was characterised by multiple, low-frequency (often singleton) haplotypes. The *CYP3A5*3* haplogroup has differentiated into more haplotypes than the remaining low/non-expresser haplogroups; although diversity is highest in the ancestral *CYP3A5*1* (expresser) haplogroup overall.

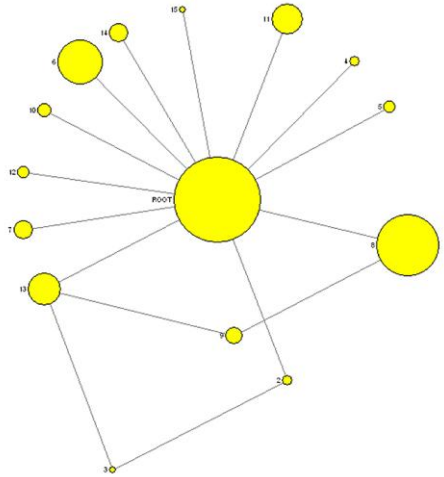
7.4.5. The distribution of microsatellite counts associated with low/non-expresser CYP3A5 alleles

A –GT microsatellite was genotyped in each of 379 Ethiopian individuals, in addition to re-sequencing of the entire *CYP3A5* coding region, exon-flanking introns and proximal promoter. *CYP3A5* haplotypes were not inferred using microsatellite data due to the wide variation in the number of repeats and the inability to accurately infer microsatellite counts onto the chromosomes of heterozygous individuals. The microsatellite repeat distribution associated with homozygotes for particular *CYP3A5* alleles is presented in Figure 7.8. The differences in the collective number of chromosomes defined by a specific variant in Figure 7.8 reflect the different number of chromosomes used for analysis of each variant. No individual was homozygous for *CYP3A5*7* and so the associated distribution of variation in –GT repeats could not be plotted for this allele.

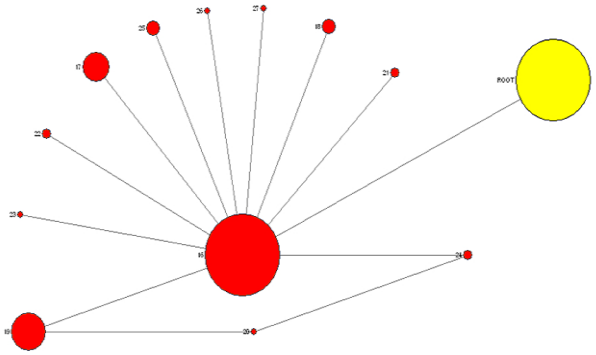
The range of microsatellite repeats associated with *CYP3A5*3* (28-39 GT repeats) is higher than that for *CYP3A5*1* (33-41) and *CYP3A5*6* (32-38)*;* although the distribution of microsatellite variation associated with the *CYP3A5*1* and *CYP3A5*6* alleles is more spread out than for *CYP3A5*3*. A striking observation about the distribution of -GT repeats in *CYP3A5*3* homozygotes is that the spread is narrow and almost all *CYP3A5*3* chromosomes have 35-36 GT repeats. Given that microsatellites normally have a high mutation rate the narrow spread of data in *CYP3A5*3* carriers is consistent with the idea that the *CYP3A5*3* allele has risen to high frequency rapidly.

Figure 7.6: Networks of a) *CYP3A5*1* haplotypes; b) *CYP3A5*3* haplotypes; c) *CYP3A5*6* haplotypes; and d) *CYP3A5*7* haplotypes inferred for a ~4kb region. Networks assume single mutational steps. The size of each circle is proportional to haplotype frequency. Haplotype codes refer to those listed in Table 7.2. "Root" refers to the ancestral sequence for the polymorphic sites; inferred from the closely related chimpanzee sequence; and identical to haplotype 1. Haplotypes defined by the *CYP3A5*3* mutation are coloured in red; those defined by *CYP3A5*6* in blue; *CYP3A5*1* in yellow; and *CYP3A5*7* in green.

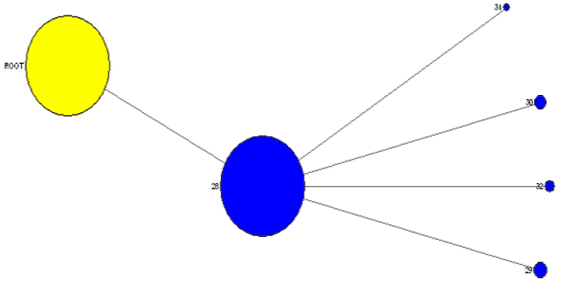
a)



b)



c)



d)

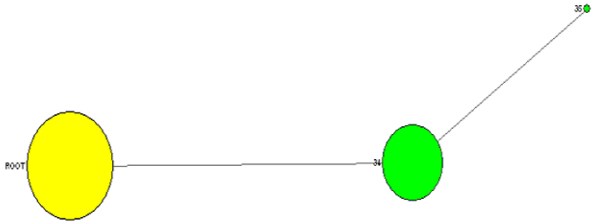
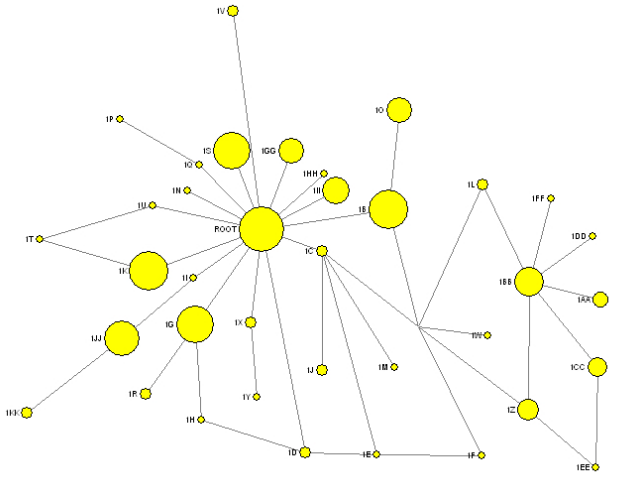
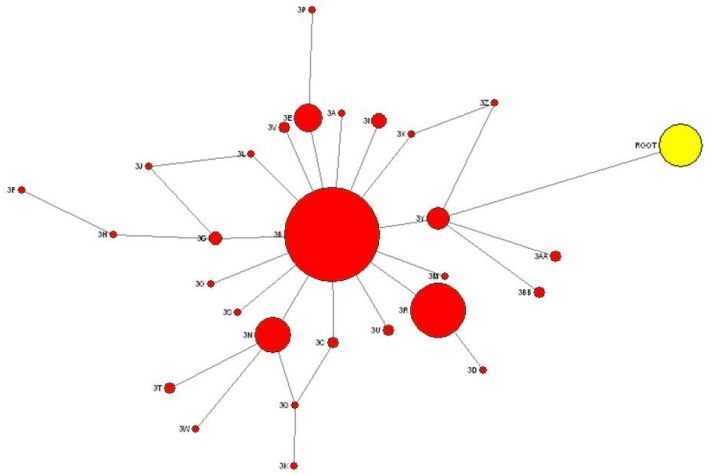


Figure 7.7: Haplotype networks of three *CYP3A5* haplogroups, inferred for an 8063bp region. Haplotypes are coloured according to the haplogroup to which they belong; *CYP3A5*3* haplotypes are shown in red; *CYP3A5*1* in yellow; *CYP3A5*7* in green; *CYP3A5*6* in blue and *CYP3A5*3/*6* recombinant haplotypes are shown in purple. The size of each haplotype is proportional to its frequency in the global database. Table 7.1 contains information on the exact composition of each coded haplotype in accordance with Figure 7.1. The haplotype named "ROOT" was used to root each network and is equivalent to the chimpanzee *CYP3A5* sequence and to haplotype 1 (Table 7.1).

*CYP3A5*1* haplogroup



*CYP3A5*3* haplogroup



*CYP3A5*6* haplogroup

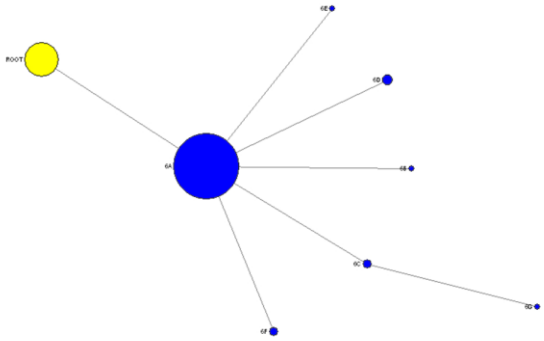
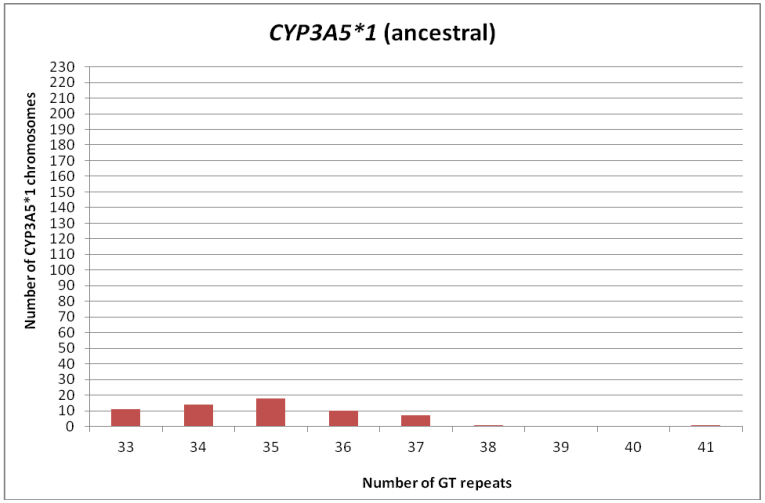
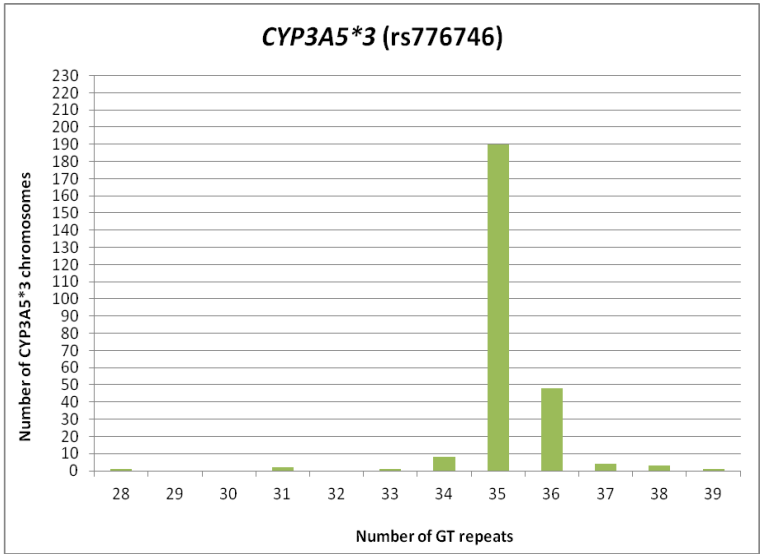
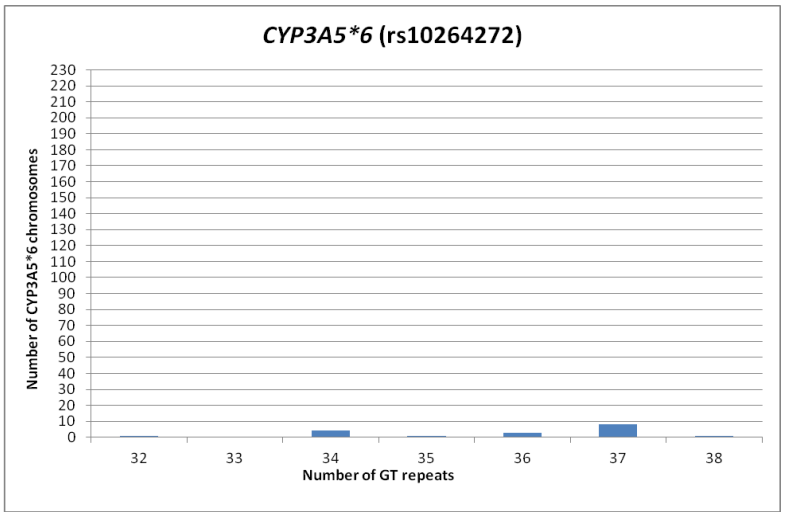
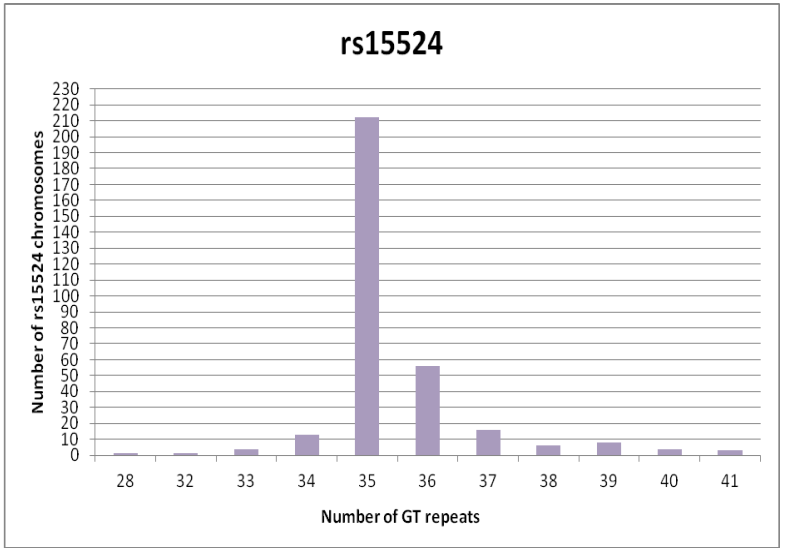


Figure 7.8: The variation in the number of –GT microsatellite repeats in individuals homozygous for one of *CYP3A5*1, CYP3A5*3, CYP3A5*6* and rs15524 alleles. The Figure shows the amount of variation in the –GT microsatellite repeats associated with each allele. Heterozygotes for one or more of the alleles were excluded from analysis.









7.4.6. Estimating the ages of common low/non-expresser CYP3A5 alleles

The age of the *CYP3A5*3, CYP3A5*6* and rs15524 variants were estimated using Ytime software (executed in Matlab, see Table 7.3). Interestingly the *CYP3A5*3* allele (~76,000 years) was estimated to be younger than *CYP3A5*6* (~200,000 years). The distribution of –GT microsatellite repeats associated with the rs15524 mutation almost completely overlaps with the *CYP3A5*3* allele (see Figure 7.8), which reflects the modal *CYP3A5* haplotype (see Table 7.1). rs15524 was estimated to be older than *CYP3A5*3,* consistent with the analysis of haplotype networks (see Figure 7.1), and suggests that the modal *CYP3A5*3* haplotype (3B) arose as a result of the *CYP3A5*3* mutation appearing on an existing haplotype with the rs15524 mutation.

An interesting finding is that the estimated age of the *CYP3A5*3* variant does not predate the approximate date for the first wave of human migration out of Africa ~100,000 years ago (Reed and Tishkoff 2006). The consensus model of human migration out of Africa estimates that ~70-80,000 years ago modern humans were in the Arabian Peninsula (Figure 7.9a) and had also migrated from East to West Africa (Figure 7.9b). The estimated age of the *CYP3A5*3* allele is consistent with its distribution in Africa (based on data generated for this thesis) and globally. An estimated young age (~76,000 years) coupled with global differences in the frequency of the allele (see chapter 3) is consistent with a hypothesis of recent spread of *CYP3A5*3* as a result of positive selection in non-African populations. However it is possible that the *CYP3A5*3* mutation is older than 76,000 years, and that selection has influenced the age estimates as a result of recent positive selection.

The distribution of the *CYP3A5*6* allele (chapter 3) is similar to that of another variant in the gene encoding the drug metabolising enzyme Flavin-containing monooxygenase 2 (FMO2); *FMO2*1*, previously genotyped in the majority of the cohort genotyped for chapter 3 (Veeramah et al. 2008). This variant is found at high frequencies throughout sub-Saharan Africa, the authors estimated the age of the *FMO2*1* variant to be 502,404 years (95% confidence intervals: 154,790 years – 1,041,243 years) based on a coalescent-based method described in (Griffiths and Marjoram 1996). The age estimates of both the *CYP3A5*6* and *FMO2*1* alleles predate the exodus of modern humans out of Africa. It is possible that the *CYP3A5*6* mutation was lost in a population bottleneck out of Africa; which would explain its African distribution and why the age estimates of this allele are older than for *CYP3A5*3.*

Table 7.3: Estimating the age of *CYP3A5* variants which define the most common haplogroups in Ethiopia. Allele ages were estimated using a mutation rate of 0.00045 (Whittaker et al. 2003) and a generation time of 32 years (Tremblay and Vezina 2000). The confidence intervals for the estimated age of the *CYP3A5*6* are large; most likely a reflection of the small sample size.

N.B. UTR is the Untranslated Region

***CYP3A5* variant**	**Location on chromosome 7**	**Location in gene**	**Allele dated**	**Number of chromosomes**	**Average squared distance (ASD)**	**Time to most recent common ancestor**		**95% confidence intervals of allele age estimate based on a star phylogeny**			
						Estimate of allele age		**Lower**		**Upper**	
						Generations	**Years**	**Generations**	**Years**	**Generations**	**Years**
CYP3A5*3	99270539	Intron 3	G	134	1.0746	2388	76,416	1797	57,504	3211	102,752
CYP3A5*6	99262835	Exon 7	A	18	3.0714	6825	218,400	3086	98.752	11975	383,200
rs15524	99245914	3´ UTR	T	324	1.8426	4095	131,040	3157	101,024	5413	173,216

Figure 7.9a: Human migratory patterns out of Africa. The dates shown are conservative estimates of when modern humans first appeared in specific geographic regions. The image has been taken from: http://ngm.nationalgeographic.com/ngm/0603/feature2/images/mp_download.2.pdf

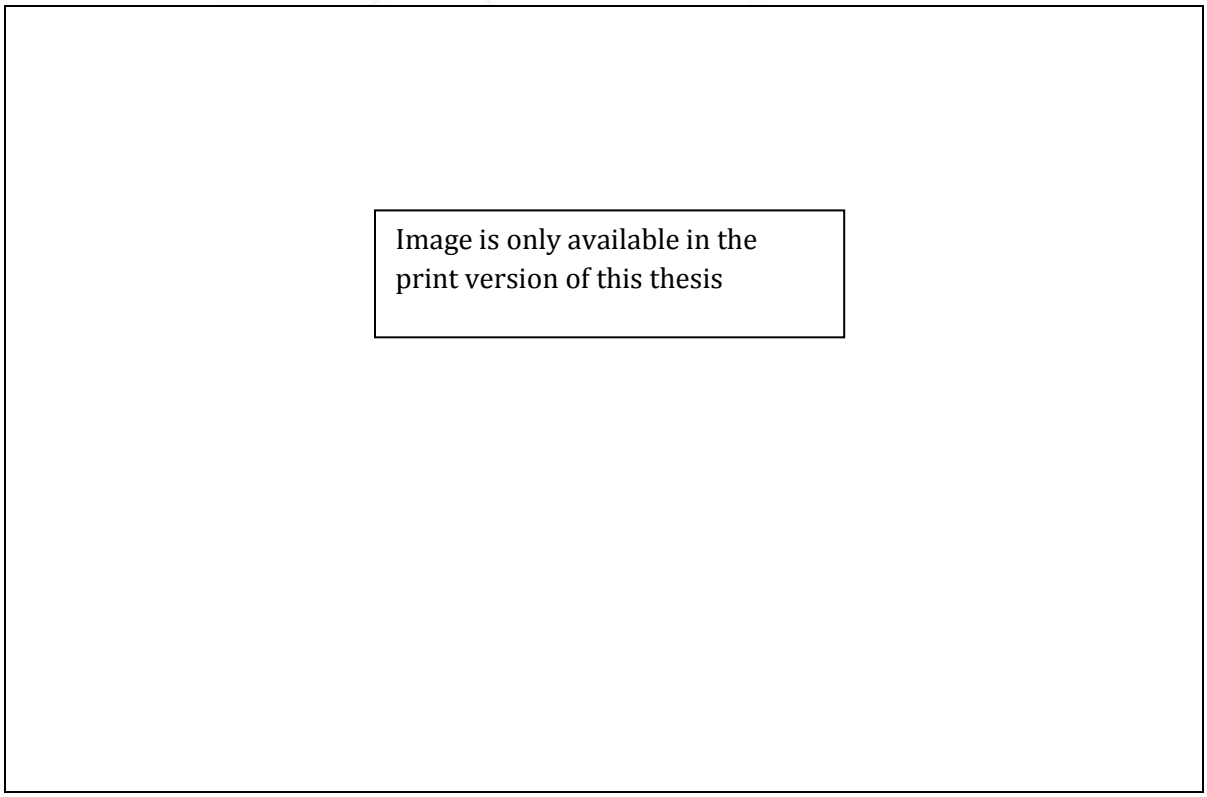
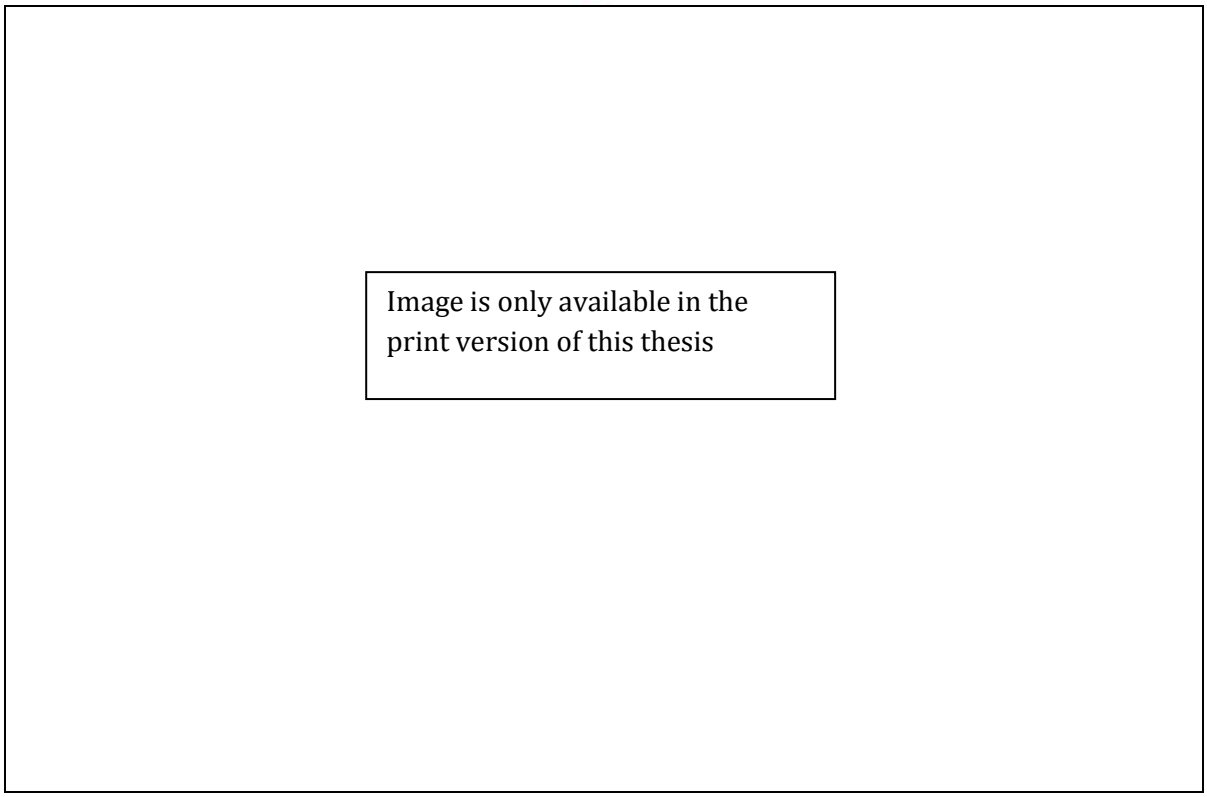


Figure 7.9b: Human migratory routes out of Africa. The dates shown are conservative estimates of when modern humans first appeared in specific geographic regions. The image has been taken from: http://datamining.typepad.com/data_mining/2009/08/the-human-journey.html



A recent study of *CYP1A2* variation in the same five Ethiopian groups re-sequenced in this thesis (Browning et al. 2010) identified a number of novel and known variants in the gene. The authors dated a number of variants from the study using microsatellite data for the populations. All variants which were observed at similar frequencies to the *CYP3A5*6* allele in Ethiopia (~20%) were estimated to be over 100,000 years old. There are reports of different selection pressures on *CYP1A2* (reported to be a target of purifying selection) (Browning et al. 2010) and *CYP3A5* (reported to be a target of positive selection for low/non-expression) (Chen et al. 2009). However the similarity in variant allele frequencies in the two genes enables a comparison of age estimates. The estimated age of *CYP3A5*6* is higher than for *CYP1A2* alleles of similar frequencies and predates the exodus of modern humans out of Africa; suggesting that the estimate for *CYP3A5*6* as being older than *CYP3A5*3* is correct.

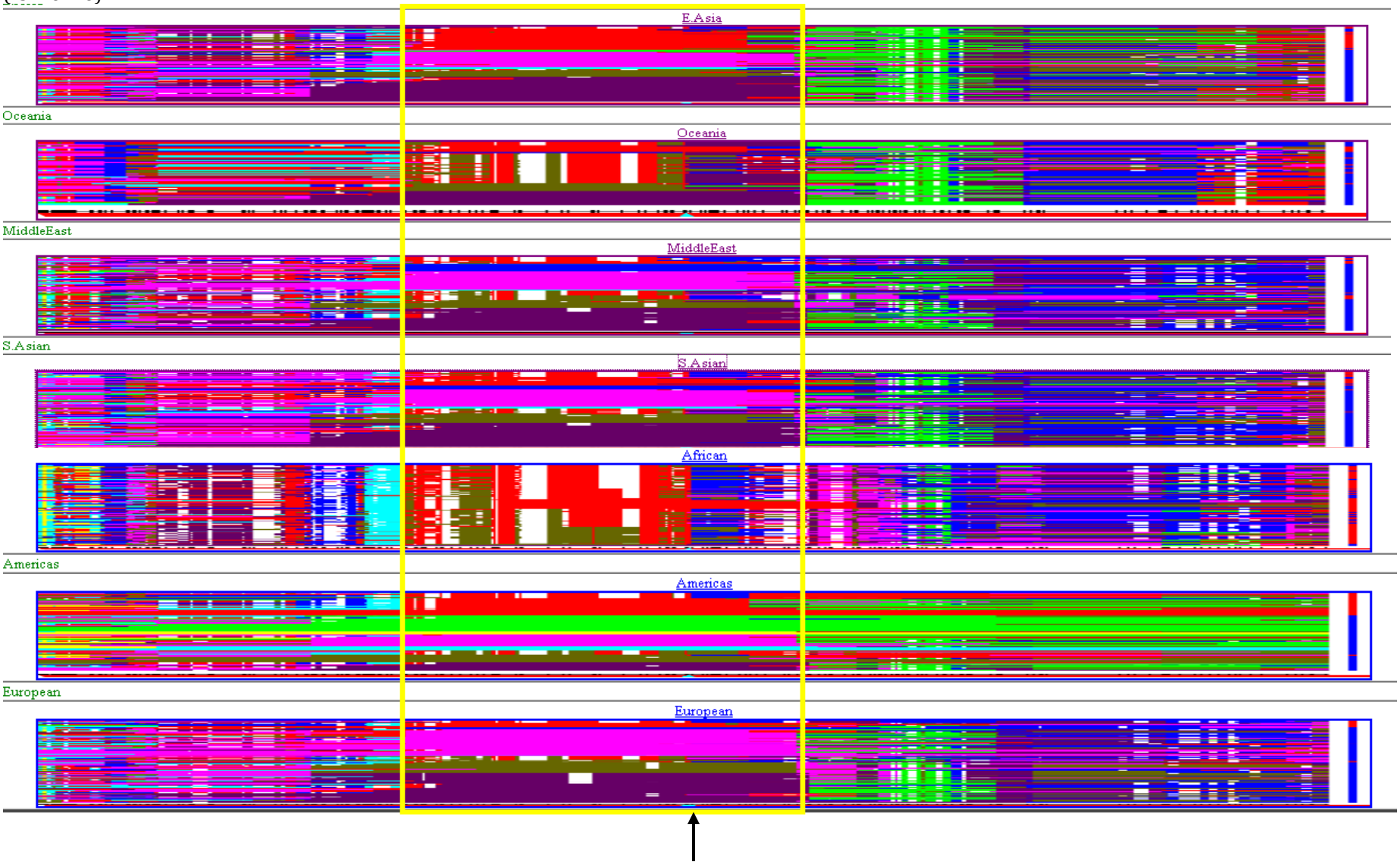
7.4.7. Examining the CYP3A5 locus for evidence of positive selection

The estimated age of *CYP3A5*3* (~76,000 years) suggests that it may have undergone a recent selective sweep, and rapid spread, in non-African populations. Its presence and global inter-population differences suggest that there are distinct environmental, or alternative, pressures, which are more pronounced as distance from the equator increases, and provide a selective advantage to having the low/non-expresser allele. Environmental pressures associated with increased distance from the equator would explain the significant positive correlation between *CYP3A5*3* allele frequencies and latitude (reported in chapter 3). This section examines evidence of positive selection on the *CYP3A5* gene in populations from HapMap phase II and the HGD panel.

7.4.7.1. Evidence from haplotype structure

Haplotypes for SNPs genotyped over a ~2Mb region of chromosome 7, including the *CYP3A5*3* mutation (rs776746), in HGDP samples were extracted from Haplotter. This allowed for an examination of a larger genomic region surrounding the *CYP3A5* gene, comparative to the Ethiopian data. The position of the *CYP3A5*3* mutation is annotated in Figure 7.10. Haplotypes were coloured according to their similarity in a method outlined by Conrad *et al* (Conrad et al. 2006). Briefly the modal, or core, haplotypes in each geographic region are identified and coloured. Each subsequent haplotype, in order of decreasing frequency, is compared against the core haplotype and coloured according to where it is similar to, and differs from, the modal. The colouring enables similarities and differences between haplotypes to be visualised (Conrad et al. 2006).

Figure 7.10: Inferred haplotypes for a 2Mb region of chromosome 7, surrounding the *CYP3A5*3* locus (rs776746), for 52 global populations from the HGDP database extracted from (http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/). Data are shown by geographic region. The region boxed in yellow shows a ~500,000 base pair region surrounding the *CYP3A5*3* mutation and *CYP3A5* gene. The arrow marks the position of the *CYP3A5*3* mutation (rs776746).



A comparison of haplotype structures by geographic region found that there is more evidence of recombination and variation on African haplotypes than those observed in other global populations (Figure 7.10). This is consistent with network analysis of haplotypes in section 7.4.3 which found more recombinant haplotypes in a ~4kb region (*n*=9) than an 8063bp region (*n*=8); and with previous reports that haplotype diversity is highest in populations with recent African ancestry (Conrad et al. 2006; Li et al. 2008; DeGiorgio et al. 2009). Haplotypes defined *CYP3A5*1* are more variable than those defined by the *CYP3A5*3* mutation; these haplotypes are also defined by more recombination events, particularly in Africans where *CYP3A5*1* is observed at high frequencies. Interestingly, there were multiple long *CYP3A5*3* haplotypes observed in global populations; suggesting that haplotypes carrying the mutation have extended homozygosity which may have arisen due to a selective sweep on *CYP3A5*3*, or another mutation in high LD with the allele. A ~500,000bp of haplotype homozygosity is annotated on Figure 7.10. The genomic sequence surrounding the annotated region is characterised by high levels of recombination and diversity in all geographic regions.

The modal *CYP3A5*3* haplotypes inferred from Ethiopian re-sequencing data (see chapter 6) are characterised by a paucity of variation, high LD between polymorphic markers and homogeneity over a ~12kb region. Additionally, there is more variation in *CYP3A5* flanking regions than in the gene itself, consistent with HGDP data. The genomic region re-sequenced in Ethiopians is much smaller than that represented by SNPs genotyped in the HGD panel and it is possible that over a large genomic region there would be more differentiation of *CYP3A5*3* haplotypes in Ethiopia than in non-African HGDP populations. Given the considerable intra-Ethiopian diversity characterised in this thesis, it is also likely that haplotypes in an equivalent region to that shown in Figure 7.10 in the Afar, Amhara and Oromo are likely to be similar to those from the Middle East; in contrast to the Anuak who are likely to have haplotypes much more similar to those observed in Africa. It is also possible that intra-African diversity within the eight non-Ethiopian African sample sets re-sequenced for this thesis (see chapter 4) will be high and further differentiate African variation at the *CYP3A5* gene.

7.4.7.2. Evidence from HapMap II and HGDP populations

As outlined in section 7.1.2 a number of different tests aim to detect signatures of positive selection in the genome. iHS scores for each HapMap II population were negative indicating that the degree of decay on derived haplotypes, carrying the *CYP3A5*3* allele, is lower (i.e. derived haplotypes are longer) than ancestral haplotypes. The greatest skew to

negative values was observed in Europeans (iHS = -1.354 for a 2Mb region surrounding the *CYP3A5*3* mutation), followed by East Asians (iHS = -0.724), and the lowest in the Yoruba (iHS = -0.589); although no estimate was at the extreme end of iHS estimates [(iHS = -2) (Voight et al. 2006)].

The extent of LD decay in a 2Mb region surrounding the *CYP3A5*3* mutation in each of three HapMap II populations is shown in Figure 7.11. Dark blue and red down the centre column of each Figure corresponds to the ancestral *CYP3A5*1* and derived *CYP3A5*3* alleles respectively. Core haplotypes defined by either allele are also coloured either dark blue or red. Genomic regions where haplotype homogeneity breaks down, for example due to recombination, are then coloured differently from the core regions. All *CYP3A5*3* haplotypes, those annotated in each of the Figures, are characterised by high LD which extends over a region of at least 0.25Mb; consistent with the haplotypes extracted for HGDP populations (Figure 7.10).

Data on the degree of haplotype decay surrounding the *CYP3A5*6* allele (rs10264272) were extracted for Yoruba individuals from the HapMap cohort (data not shown). The *CYP3A5*6* locus is monomorphic in Europeans and East Asian populations and iHS scores for the region surrounding the *CYP3A5*6* allele could not be obtained for these populations. The iHS score for the region immediately surrounding *CYP3A5*6* (iHS = -1.351) suggests that haplotype homozygosity surrounding the *CYP3A5*6* mutation is higher than that surrounding *CYP3A5*3* in the Yoruba.

Integrated haplotype scores (iHS), which were calculated by (Li et al. 2008), for a 2Mb region surrounding the *CYP3A5*3* mutation were extracted from Haplotter (see Figure 7.12). Differences in iHS estimates between geographic regions were observed. Across the 2Mb region iHS estimates were highest in populations from Europe, Central Asia, East Asia and the Middle East. iHS values that are ≥2 are considered to provide evidence of positive selection (Voight et al. 2006). iHS scores indicative of positive selection were identified in populations from the same geographic regions as those reported in which there is strong evidence that the *CYP3A5*3* allele has undergone a selective sweep (Li et al. 2011). The iHS scores, coupled with the haplotype structure observed in Figure 7.10 is consistent with the hypothesis of a selective sweep having acted on the *CYP3A5*3* outside of Africa (Thompson et al. 2004), and that the allele is advantageous outside of the African continent.

Figure 7.11: The decay of haplotypes over a 2Mb (2,000,000bp) region surrounding the *CYP3A5*3* allele compared to the ancestral *CYP3A5*1* allele. Core haplotypes defined by either the *CYP3A5*1* or *CYP3A5*3* allele are coloured dark blue and red respectively. Genomic regions where haplotype homogeneity breaks down, for example due to recombination, are then coloured differently from the core regions. Horizontal lines are haplotypes; SNP positions are shown below each haplotype plot. The regions within dark blue boxes in each Figure highlight the *CYP3A5*3* haplotypes in each population. Plots are shown for 60 individuals of northern and western European ancestry (CEU); 89 Japanese and Han Chinese individuals from Tokyo and Beijing respectively (ASN); and 60 Yoruba from Ibadan, Nigeria (YRI). The images have been taken from the Haplotter website (http://haplotter.uchicago.edu/)

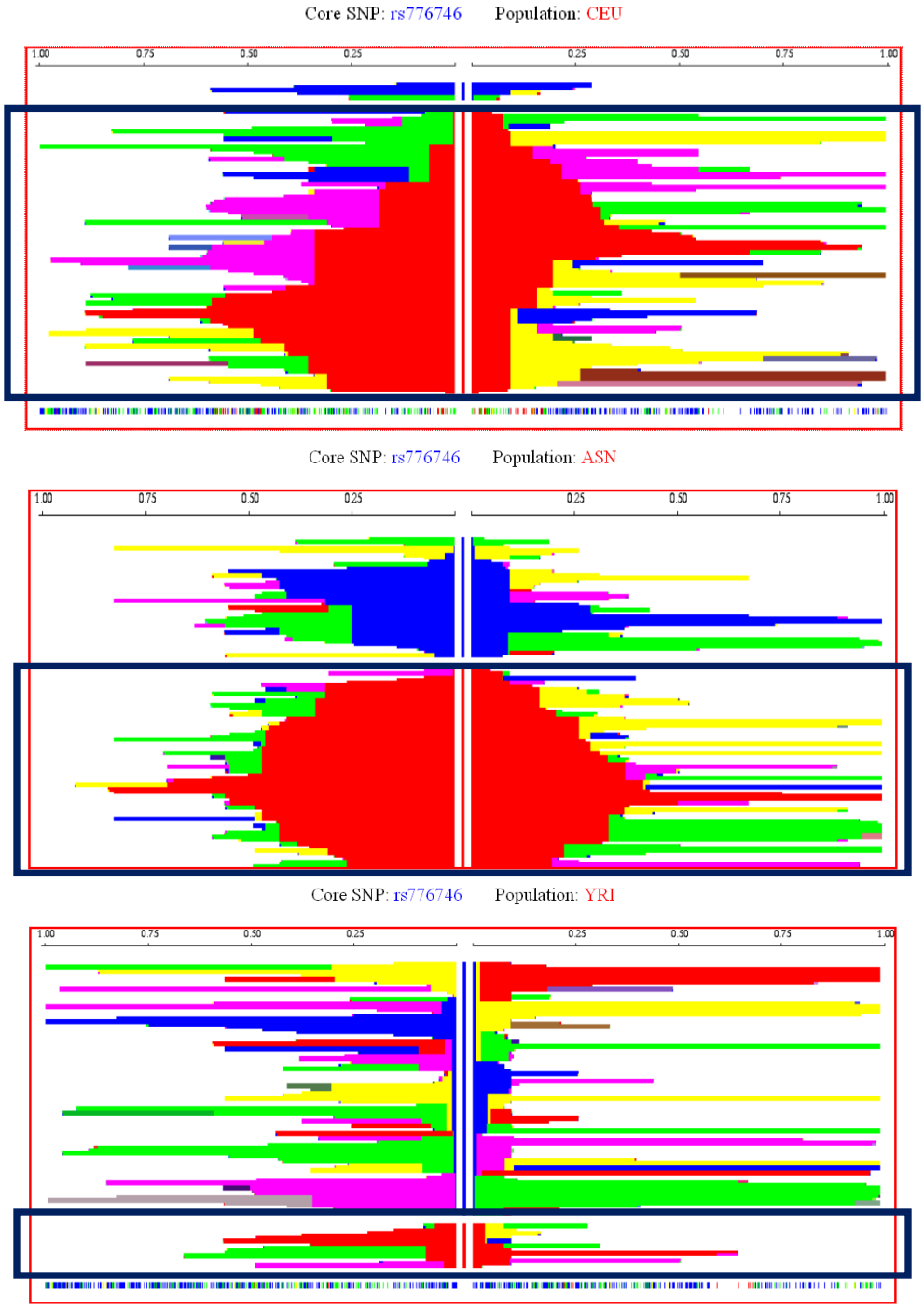
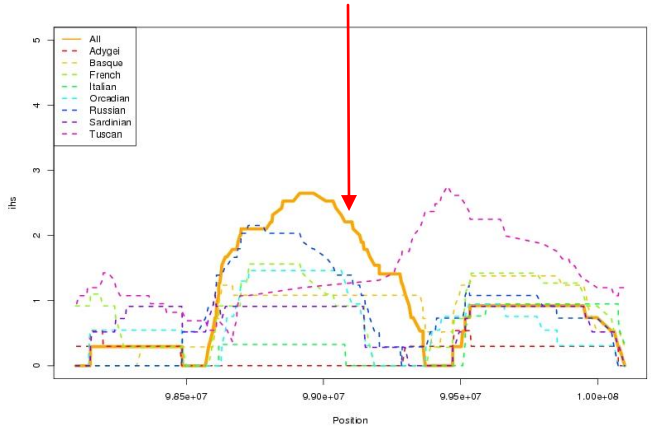
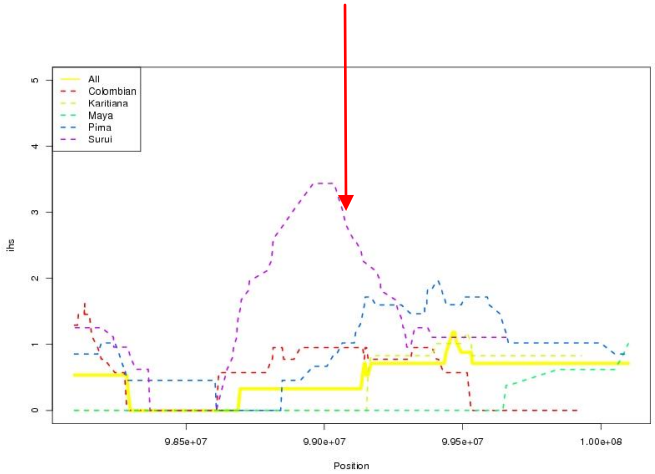


Figure 7.12: integrated haplotype score (iHS) for populations from seven geographic regions from the HGD panel. Log transformed data are presented for a 2Mb region of chromosome 7 which incorporates the *CYP3A5* gene region. Data have been extracted from the Haplotter website (http://haplotter.uchicago.edu/) (Voight et al. 2006). In each Figure, the red arrow indicates the position of the *CYP3A5*1/*3* locus.

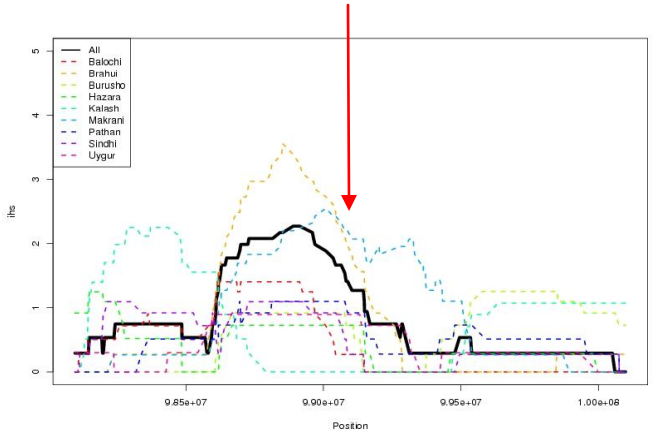
Europe



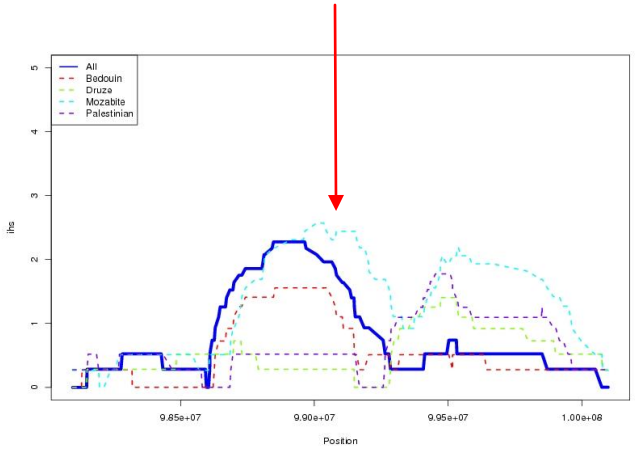
Americas



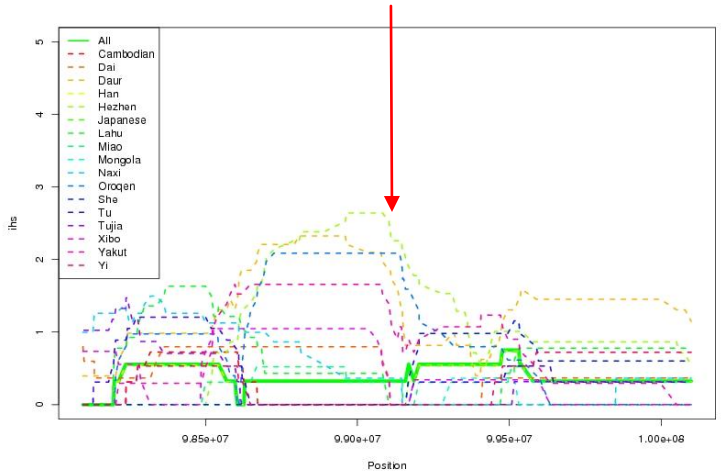
Central Asia



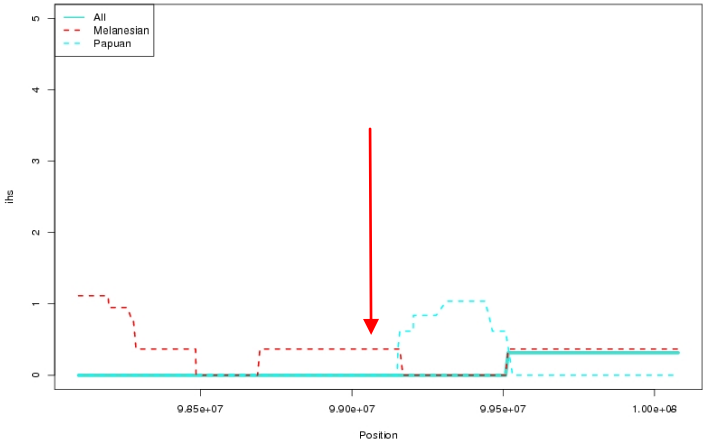
Middle East



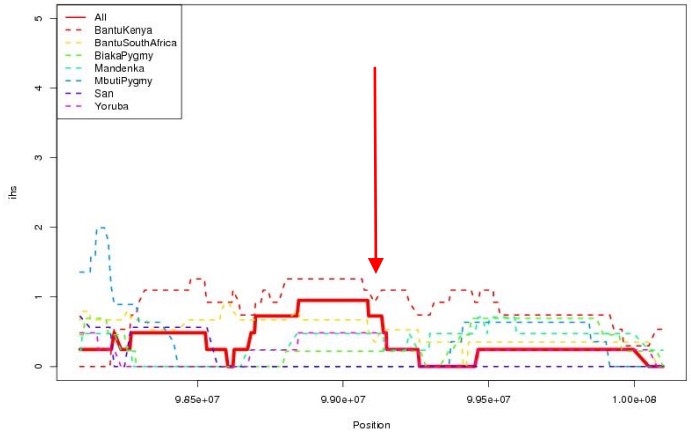
East Asia



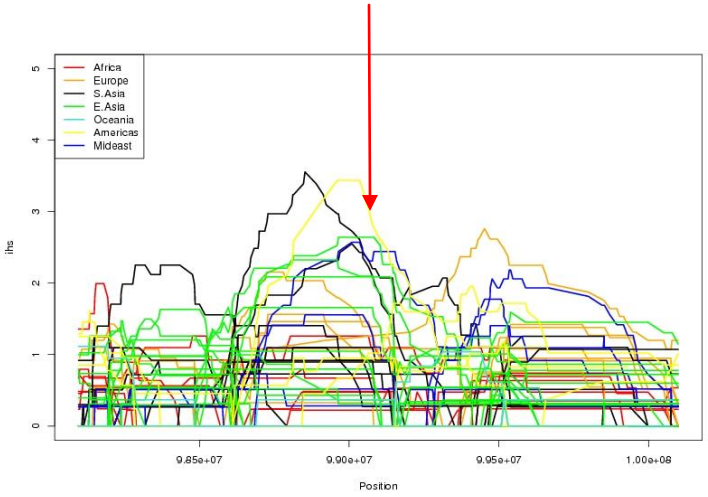
Oceania



Africa



All geographic regions



7.4.7.3. Examining signatures of selection at the CYP3A5 locus in the context of other known targets of selection

Genomic scans for positive selection have identified additional genes on chromosome 7 which are believed to have undergone positive selection; such as *CD36* (Sabeti et al. 2006). *CD36* encodes a cell surface receptor protein which is involved in multiple physiological processes including angiogenesis, thrombosis and atherogenesis (Martin et al. 2011). Mutations in *CD36* are protective against cerebral malaria caused by the parasitic isoform *Plasmodium falciparum* (Aitman et al. 2000; Pain et al. 2001). Erythrocytes infected with red blood cells often aggregate and can cause some of the more severe pathological effects of malaria such as vasocclusion (the restriction of blood flow through the vessels). CD36 proteins mediate adhesion of infected erythrocytes (Cserti-Gazdewich et al. 2011), similarly to the ABO blood group; where the A sugar mediates adhesion of infected red blood cells but the O group does not; and is found at high frequencies in malaria endemic regions (Cserti-Gazdewich et al. 2011). Mutations in the *CD36* gene are highly prevalent in *Plasmodium falciparum* endemic regions and are associated with reduced adhesion of infected red blood cells; so reducing malaria-associated pathological effects (Aitman et al. 2000).

Given that *CD36* is known to have been the target of positive selection in populations exposed to *Plasmodium falciparum* malaria (predominantly in Africa); iHS estimates for a 2Mb region immediately surrounding the *CD36* gene were extracted for African populations from HapMap (Figure 7.13) and the HGD panel (Figure 7.14). African populations from both datasets had significantly higher iHS scores than non-African populations ($p \leq 0.001$). The iHS scores reported for *CD36* are similar to those reported for *CYP3A5* in non-African populations. A comparison of *CYP3A5* iHS estimates with those for the gene encoding the enzyme lactase, which has been shown to have undergone positive selection in Europe, (Tishkoff et al. 2007; Ingram et al. 2009; Gerbault et al. 2011), were similar (data not shown). The data that are available for HapMap and HGDP-CEPH populations provide support for a hypothesis of positive selection on the *CYP3A5*3* allele in non-African populations.

Figure 7.13: comparisons of iHS estimates for *CD36* in populations from the HapMap II panel. Note the large iHS values obtained for the Yoruba comparative to the European and East Asian populations. Data have been extracted from the Haplotter website (http://haplotter.uchicago.edu/) (Voight et al. 2006). The serrated line shows the position of the *CD36* gene where there is an increase in the iHS score.

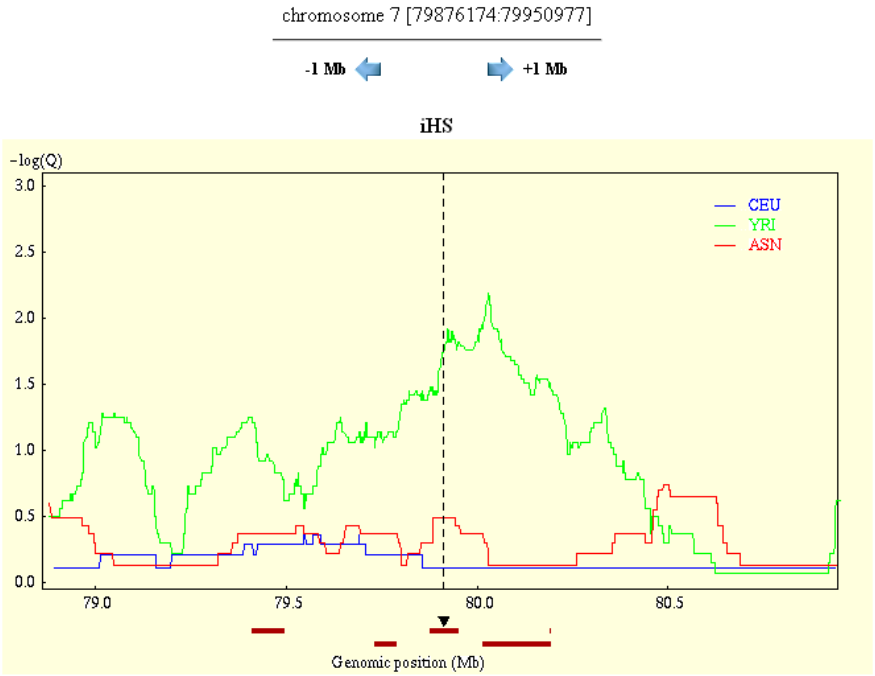
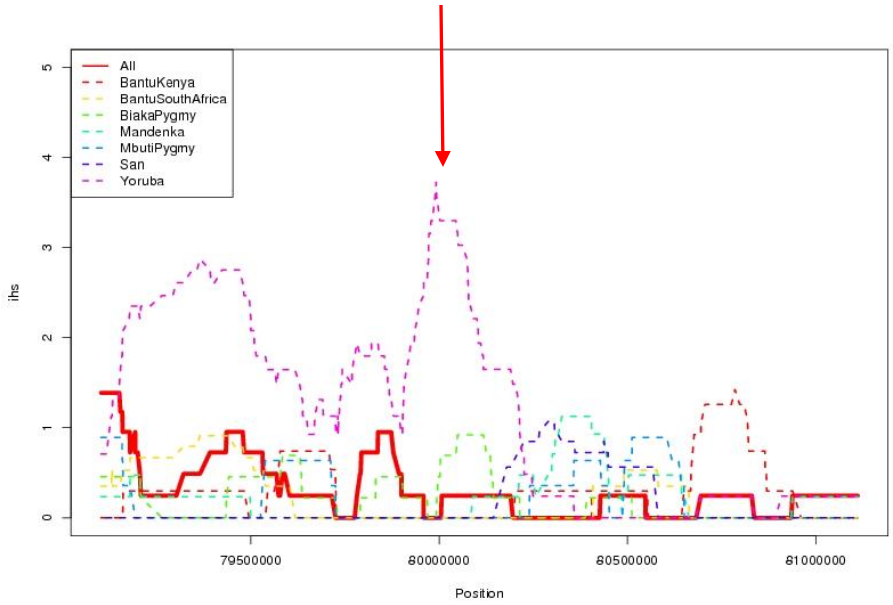


Figure 7.14: iHS scores for a 2Mb region of chromosome 7 surrounding the *CD36* gene (annotated on the Figure). There is a significant increase in the iHS score in the region surrounding *CD36*. Data are presented for the African populations which are part of the HGD panel, and have been extracted from the Haplotter website (http://haplotter.uchicago.edu/) (Voight et al. 2006).



7.5. Discussion

7.5.1. The evolutionary relationships between CYP3A5 haplotypes are characteristic of rapid growth

Haplotype networks were constructed to show the most parsimonious relationships between *CYP3A5* haplotypes inferred from the re-sequencing data. A star phylogeny was inferred from all global haplotypes; characteristic of rapid population growth or a selective sweep (Martins and Housworth 2002). Differentiation within *CYP3A5* haplogroups is predominantly a result of novel mutations arising on a single modal haplotype within each haplogroup. Although there is evidence of recombination across the haplotype networks; particularly in populations with recent African ancestry.

Despite the lower age estimate of the *CYP3A5*3* allele (~76,000 years), comparative to *CYP3A5*6* (~200,000 years), greater differentiation of *CYP3A5*3* haplotypes was observed. There is strong evidence of positive selection acting on the *CYP3A5*3* allele in populations outside of sub-Saharan Africa (discussed further in section 7.5.3). Rapid growth and a global prevalence of the allele are consistent with the inferred star-like phylogenies. Selection and rapid growth would explain one modal *CYP3A5*3* haplotype with multiple, low frequency haplotypes in the haplogroup. This is seen in all populations with high frequencies of the *CYP3A5*3* allele.

The ancestral haplotype, named "ROOT", which is observed at low frequency in the re-sequenced cohort and is identical to the chimpanzee sequence for the identified polymorphic sites, was manually used to the root the haplotype networks. Interestingly in Ethiopia no individual was observed to be homozygous for this haplotype; not even in the Anuak who have the highest frequencies of the *CYP3A5*1* allele (see chapter 3). From re-sequencing data for the entire gene, all individuals in every population have at least one haplotype with a derived allele; and the modal haplotype was defined by the *CYP3A5*3* mutation. As discussed in chapter 6, conserved genes tend to have low tolerance for damaging mutations, which are removed by purifying selection, unlike pseudogenes or those which are targets of positive selection (Graur et al. 2000). This is consistent with directional, or positive, selection for low/non-expression of CYP3A5 protein.

*7.5.2. The CYP3A5*3 mutation is estimated to have arisen after the exodus of modern humans from Africa ~100,000 years ago*

Fossil and genetic diversity data support an African origin of modern humans ~200,000 years ago (White et al. 2003; McDougall et al. 2005; Campbell and Tishkoff 2008). The recent African origin model of human evolution estimates that the first exodus of modern humans out of the continent, via East Africa, occurred ~100,000 years ago (Tishkoff and Verrelli 2003) initially into the Middle East and West Africa 70-80,000 years ago; followed by subsequent expansions into other geographic regions (see Figures 7.9a-b).

The *CYP3A5*3* mutation is estimated to have arisen ~76,000 years ago, based on data within this chapter. The age of the mutation is consistent with its presence within Africa as well as in other global populations. It is possible that the allele is older than the estimates within this thesis; the microsatellite mutation rate within the genomic region may differ from the genome average for dinucleotide repeats ($4.5x10^{-4}$) used in this chapter, and this may have influenced estimates generated from the stepwise mutation model. If selection has played a role in driving the allele to high frequency in populations outside of Africa then this would also influence the allele age estimates and predict them to be younger than they really are. Positive selection would drive *CYP3A5*3* and all tightly variation, including microsatellite markers, to high frequency at a rate that mutation would be unable to generate diversity in the number of repeats. Decreased diversity in microsatellite repeat numbers would influence estimates of the allele age.

The estimated age of *CYP3A5*3*, coupled with a significant increase in its frequency in populations outside of Africa suggest that the allele may be advantageous outside of the continent. The frequencies of *CYP3A5*3* are high in non-African populations; almost to fixation in some European groups. The estimated age of *CYP3A5*3* show that it is not a recent mutation onto a haplotype background with a paucity of additional variation, which means that it is a plausible candidate on which positive selection could have acted (see section 7.5.3). An age of ~76,000 years is consistent with its presence within and outside of Africa. The shift in allele frequencies, coupled with other selection signatures on haplotypes defined by this allele, in non-Africans suggest that the allele may have undergone selection when modern humans first appeared in the Middle East due to an environmental change causing it to be advantageous outside of Africa. This hypothesis would be consistent with the observed positive correlations between *CYP3A5*3* frequencies and increased latitude (reported in chapter 3). It is plausible that an increase in latitude of ~20° (difference between the Middle East and Ethiopia; the likely exit point for modern humans out of Africa) is coupled with

specific environmental changes in temperature and precipitation which provided an initial selective pressure which was then further exacerbated as humans migrated to higher latitudes; in Europe and East Asia. For example across populations from the HGDP-CEPH panel, (Young et al. 2005) observed that 74% of all variants observed in populations within 10° of the equator were a result of heat adaptation. This provides support for the hypothesis of strong selective pressure for low/non expression of CYP3A5 at latitudes ≥20°. There are latitudinal correlations observed within Africa as well; and frequencies of the *CYP3A5*3* allele are higher within Southern Africa than elsewhere; although an examination with heterogeneous Southern African groups will reveal whether there are selective pressures on the allele within the continent as well.

The age of the rs15524 mutation was estimated as it is found on the modal *CYP3A5*3* haplotype background in Ethiopia (see Table 7.1). The age of this allele (~131,000 years) is estimated to be older than *CYP3A5*3* which is consistent with network analysis of haplotypes and suggest that *CYP3A5*3* evolved onto an existing haplotype background with the rs15524 mutation.

The estimated age of the *CYP3A5*6* mutation (~214,000 years) predates estimates of the exodus of modern humans from Africa. Despite a paucity of variation observed on *CYP3A5*6* haplotype backgrounds (see chapter 6), the allele age is estimated to be quite old. The global distribution of the allele is almost entirely restricted to sub-Saharan Africa; with the exception of low frequencies observed in North Africa and in the Arabian Peninsula. These observed frequencies are almost certainly a result of population admixture with East African populations. The age, and restricted geographic distribution, of the allele suggest that it was lost in a population bottleneck when modern humans first left Africa. The allele frequency across sub-Saharan African populations is very similar (18-20%); consistent with its spread being related to intra-African migrations. Additionally, an examination of inferred Ethiopian *CYP3A5*6* haplotypes for a 12,237 base pair *CYP3A5* region (see chapter 6) found that *CYP3A5*6* is in high LD with other derived variants. The results of Fay and Wu's *H* test of high frequency derived variants were not significant; although this is likely to be due to a paucity of variation observed within the *CYP3A5* gene. However high frequency derived variants, on existing *CYP3A5*6* haplotype backgrounds and a paucity of variation on haplotypes suggest that the allele may have been a target of selection, although not necessarily recent selection (i.e. not within the last ~30,000 years and detected by the LRH test), within Africa (Sabeti et al. 2006).

The age of the *CYP3A5*7* mutation could not be estimated as there were no observed homozygotes, for this allele, within the Ethiopian cohort. However, the distribution of the

allele is almost exclusively in Niger-Congo speaking population (chapter 3). This suggests that it has an origin that is much more recent than either *CYP3A5*3* or *CYP3A5*6,* and that its distribution within sub-Saharan Africa can be explained by recent migrations across the continent ~4000 years ago.

*7.5.3. There is evidence of positive selection acting on the CYP3A5*3 allele in populations outside of Africa*

In chapter 6, it was reported that there is a skew towards rare variants in all 8 populations from the Coriell Repositories and Ethiopia. The estimated ages of common low/non-expresser *CYP3A5* alleles (reported in this chapter) were consistent with them being old enough to have been true targets of selection. These findings are consistent with a hypothesis of positive selection having acted on the *CYP3A5* gene in non-African populations. This chapter aimed to examine these selection signatures further in populations from HapMap phase II and HGDP-CEPH.

Data extracted from online resources were consistent with a hypothesis of positive selection driving the *CYP3A5*3* allele to high frequency; iHS estimates found that the haplotypes surrounding the *CYP3A5*3* mutation extend ~500kb around the allele in non-African or Oceanian populations. Haplotypes extracted from online databases found that those carrying the derived *CYP3A5*3* allele have extended regions of homozygosity than those carrying the ancestral *CYP3A5*1* allele. Crucially, the observed iHS scores for *CYP3A5*3* haplotypes were similar to those for genomic regions believed to have undergone positive selection. Additionally, iHS scores which are considered evidence of positive selection were observed in populations from geographic regions where the *CYP3A5*3* allele is believed to have been the target of a selective sweep (Li et al. 2011). The results from (Li et al. 2011) were consistent with those reported previously (Thompson et al. 2004; Thompson et al. 2006); the allele frequency spectrum for *CYP3A5* shows strong evidence of a selective sweep. The data presented here for HapMap phase II and HGDP populations extend the findings by examining haplotypes around a large genomic region surrounding the *CYP3A5*3* mutation. The data provide further evidence of positive selection on the low/non-expresser variant; and that the allele is advantageous outside of Africa.

Although there were not enough Ethiopian data to obtain iHS estimates, a comparison of the haplotype structure within these populations with groups from the HGDP cohort found that there were similarities in haplotype structure and recombination patterns between Ethiopians and populations from the Middle East. It has been previously mentioned, that low

frequencies of the *CYP3A5*6* allele observed in the Middle East are likely to be due to gene flow between Ethiopia and populations from the Arabian Peninsula. Additionally, there is a known Arabian contribution to Ethiopian ancestry (Kivisild et al. 2004; Lovell et al. 2005) which is a likely contributor to the observed diversity seen in *CYP3A5* haplotypes in the Afar, Amhara and Oromo.

If a specific environmental variable(s) associated with increased latitude was a selective factor which originally drove the *CYP3A5*3* allele to high frequency then its frequency in the Afar, Amhara and Oromo is anomalous; given that these groups are on almost the same latitudinal plane as West and West Central African groups (which have lower *CYP3A5*3* allele frequencies, see chapter 3). It is unlikely that selection would have caused high frequencies of *CYP3A5*3* within Ethiopia. The effect that such high *CYP3A5*3* allelic frequencies may have on Ethiopian healthcare is discussed in more detail in chapters 4 and 6.

An LRH test, with larger genomic regions re-sequenced or genotyped in Ethiopian populations, would aid in elucidating how far haplotype homogeneity extends within these populations. It is possible that over longer genomic regions; haplotype homogeneity would break down. It is also possible that, as LD extends across large genomic regions of chromosome 7; in all global populations that have been genotyped and re-sequenced, that *CYP3A5*3* is not itself a target of selection but instead in high LD with a selection target located further along chromosome 7. Given the paucity of Ethiopian data generated for this thesis this cannot be ruled out.

A strong positive latitudinal correlation is observed for genes that are involved in the regulation of hypertension (Young et al. 2005). If *CYP3A5* is involved in the regulation of and differential susceptibilities to salt-sensitive hypertension, as is suggested by a strong positive correlation between the *CYP3A5*3* allele and latitude, then this could be an explanation for why the allele is likely to have undergone a selective sweep. A correlation between *CYP3A5*3* allele frequencies and latitude was detected using Spearmann's Rank Correlation analysis (chapter 3). An alternative to see how latitude affects *CYP3A5*3* allele frequencies could be modelled using genetic boundary analysis, as described by (Barbujani et al. 1989); whereby the effect of an increase in latitude of e.g. 0.5° is used to estimate the corresponding allele frequencies based on the observed data. This would map the effect of latitude on *CYP3A5*3* allele frequencies at a fine scale and elucidate how much of an effect this environmental factor has on *CYP3A5*3* allele frequencies.

7.5.4. Potential further analyses of positive selection at the CYP3A5 locus using simulated datasets

Phylogenetic trees/networks enable the processes by which haplotype diversity arose to be mapped; dating of specific alleles allows old and new alleles to be differentiated; and tests such as LRH, iHS and ALnLH (which measures FRC) analyse patterns in particular datasets and examine whether they are characteristic of positive selection signals.

Another method to detect evidence of selection is to compare real and simulated datasets to see whether haplotype composition and diversity observed is indicative of particular selective patterns and constraints. *In silico* simulations enable population geneticists to generate datasets, which are as close to the real data as possible. Simulated datasets are useful tools in population genetics as all factors which have lead to a particular patterns within the dataset are known (Ritchie and Bush 2010). For example, for each individual analysed within this thesis information on the number of polymorphic sites, the statistical associations of each polymorphic locus (haplotypes and LD) and the diversity of the population from which each individual has been sampled, is known. It is therefore possible to model patterns of diversity and haplotype structure at the *CYP3A5* locus as a result of selection, genetic drift and population bottlenecks; and compare what is seen *in silico* to what is observed in the African datasets.

In silico analysis would aid in differentiating between true and mimicked selection signatures by controlling for environmental and demographic factors which can shape diversity in a particular gene. They would be useful here for the Ethiopian data in particular to ascertain whether haplotype patterns and diversity observed within certain populations from the Ethiopian cohort are likely to have arisen due to Arabian admixture or are true selection signals.

8. General Discussion

8.1. A review of the main findings of this thesis

The overall aims of this thesis were to characterise human genetic variation in the *CYP3A5* gene in geographically, linguistically and ethnically distinct African populations; to identify the potential healthcare implications of *CYP3A5* variability; and to examine evidence of positive selection acting on the gene. In summary the main findings were:

1. There is considerable inter-population variation in the *CYP3A5* gene within Africa; recent demographic history is likely to have shaped African differences at the *CYP3A5* locus.

2. The low/non-expresser *CYP3A5*3* allele (rs776746) has a global distribution; and is found at high frequencies in some East African groups. The low/non-expresser *CYP3A5*6* (rs10264272) and *CYP3A5*7* (rs41303343) alleles are also found at appreciable frequencies within Africa; indicating that multiple variants may be affecting protein expression levels within the region.

3. There is a strong positive correlation between latitude and *CYP3A5*3* allele frequencies; consistent with a hypothesis that this variant is protective against salt-sensitive hypertension.

4. There are additional, previously unidentified, variants which may be affecting expression of CYP3A5 protein within Africa.

5. The estimated age of the *CYP3A5*3* allele (~76,000 years old) and its global distribution suggests that the allele has increased in frequency recently, and rapidly outside of Africa. The age of the *CYP3A5*6* allele (~200,000 years), and its restricted geographic distribution to the African continent, suggest that the allele was lost in a population bottleneck when modern humans first left Africa ~100,000 years ago.

This chapter discusses the implications of the main results of this thesis by placing the findings in the context of current research; areas for further research will be then identified in chapter 9.

8.2. CYP3A5 variability in Africa

The out of Africa model of modern human origins estimates that modern humans evolved in Africa ~200,000 years ago and first left the continent ~100,000 years ago. Genetic diversity is comparatively higher in African populations than in other global groups; consistent with a recent African origin of modern humans (Tishkoff et al. 2009; Campbell and Tishkoff 2010). Consistent with other genomic studies, including those on genes which encode important drug metabolising enzymes (Browning et al. 2010), *CYP3A5* variability is highest in populations with recent African ancestry.

Considerable heterogeneity was observed in East Africa and differences between Ethiopian groups and others from the continent were observed. A previous study on *CYP3A5* in Ethiopia did not account for inter-ethnic variability in the individuals that were sampled (Gebeyehu et al. 2011). This thesis found considerable inter-population differences within Ethiopia. This is consistent with previous studies on Ethiopian populations which have examined genetic differences by ethnicity (Browning et al. 2010), (*CYP3A4* unpublished data, Creemer O. *et. al*) and Ethiopian Y-chromosome data (Plaster C., *et. al*, unpublished data).

CYP3A5 haplotype structure and allele frequencies in the Afar, Amhara and Oromo, all Afro-Asiatic speaking groups, are characteristic of non sub-Saharan African populations and of what would be observed after recent positive selection (Biswas and Akey 2006; Sabeti et al. 2007). There is an excess of rare variants and low/non-expresser haplotypes have a paucity of variation in addition to low/non-expresser mutations. The Afar, Amhara and Oromo ethnic groups are found in the North East of Ethiopia and of all five groups re-sequenced for this thesis are geographically closest to the Arabian Peninsula (Henze 2000). There is a known Arabian contribution to Ethiopian ancestry as a result of back migration of Semitic speaking groups into the region ~5000 years ago (Cruciani et al. 2002; Kivisild et al. 2004) and as a result Ethiopian genetic diversity is intermediate between sub-Saharan African and Eurasian groups (Lovell et al. 2005). The back migration of Semitic speakers into Ethiopia occurred at approximately the same time as the first long range migration of Bantu-speaking agriculturalists across Africa (~4000 years ago). These migrations occurred relatively recently, in the context of human evolutionary history. It has been proposed that the positive correlation between increased latitude and *CYP3A5*3* allele frequencies are beneficial in populations closest to the equator. From this explanation the high *CYP3A5*3* frequencies in three of the Ethiopian groups are anomalous. It is possible that there has been an alternative selective pressure on the *CYP3A5* gene in Ethiopia. However the most likely explanation for

the patterns of *CYP3A5* diversity observed within Ethiopia is due to extensive Arabian admixture in the three North Eastern groups compared to the Anuak and the Maale.

8.3. There are potential medical implications for African populations due to CYP3A5 variability

CYP3A5 is involved in the metabolism of many clinically used drugs which are in widespread use for diseases prevalent in sub-Saharan Africa (Lamba et al. 2002; Wojnowski 2004). Variability in protein expression is associated with differential efficacy of drug treatment and susceptibility to diseases including hypertension (Givens et al. 2003; Bochud et al. 2009) and numerous cancers (Dandara et al. 2005; Zhenhua et al. 2005).

The data generated for this thesis, particularly in chapters 3 and 4, show that there is likely to be high levels of intra-African variability in protein expression. The study of intra-African variability in medically important genes would be of benefit for the world's most vulnerable populations, and for the African Diaspora. The quantity and quality of re-sequencing data that are available on public databases is expanding rapidly with the availability of new and cheaper sequencing technologies; in a so-called "genomics revolution" (Butler 2001; Friedrich 2001; Chung 2010). Until recently, few African genomic data were available on public databases. However a number of African populations have been included for re-sequencing as part of the 1000 Genomes project (Kaiser 2008). Although the 1000 Genomes data are currently low coverage, inclusion of multiple African groups presents a way forward for genomics and medical research. Additionally, as re-sequencing technologies become cheaper, with the progression towards the $1000 genome (Bennett et al. 2005; Mardis 2006; Service 2006; Wade 2006; Defrancesco 2012), it is possible to sequence more genomes of the world's most diverse, unique and vulnerable populations to aid in disease research and the tailoring of medical treatments worldwide.

Advances in genomics are improving the availability of re-sequencing technologies and data. Currently these approaches are not yet widely available and instead approaches focusing on population based variability and diversity in medically important genes are good strategies. Over 90% of the global disease burden is found in developing countries and a substantial number of developing countries are within Africa (Aspray et al. 1998). The inclusion of ethnically and geographically diverse African populations in large scale re-sequencing efforts would aid in focusing population-specific medical intervention efforts across the continent; as mass drug administration administers drugs at dosages that were optimised for patients with recent European ancestry. It is possible that re-sequencing technologies may be unattainable, even at $1000 for whole genome re-sequencing, in many

developing countries. Therefore population genetics based approaches whereby dosages are optimised based on population-specific genetic variability are ideal. At present there are not enough data on the associated adverse clinical outcomes with mass drug intervention within Africa and therefore it is not known how European centric dosages are influencing therapeutic outcomes within the region.

So what approaches are there to study or tailor medical treatment based on specific population variability? One strategy is that described within chapter 3, where clinically relevant variants, of a medically important gene, are genotyped in multiple populations from across continental regions. This would provide an overview of genetic diversity levels within a particular geographic region and aid in understanding where specific drug regimens are likely to need adjusting. The data presented in chapter 3 suggest that for CYP3A5 substrates, East Africans are likely to be split into those populations who require non-African dosages of CYP3A5 drug substrates, and those which do. Further analysis of unique populations by re-sequencing of medically important genes will also provide an overview of whether there are novel and rare variants, in addition to well characterised ones, which may also influence protein expression and therapeutic outcomes. Of course drug metabolism is often complex with multiple enzymes influencing the metabolism of a specific drug. An example is the well studied immunosuppressant tacrolimus; CYP3A5 is one of multiple enzymes involved in its metabolism (Hesselink et al. 2003; Zheng et al. 2003). However variability in CYP3A5 expression has been shown to directly affect therapeutic outcomes associated with tacrolimus treatment. Ideally analyses of genetic variability in multiple medically important genes in specific populations will be a progression in medicine; however single gene studies are also useful in predicting clinical outcomes.

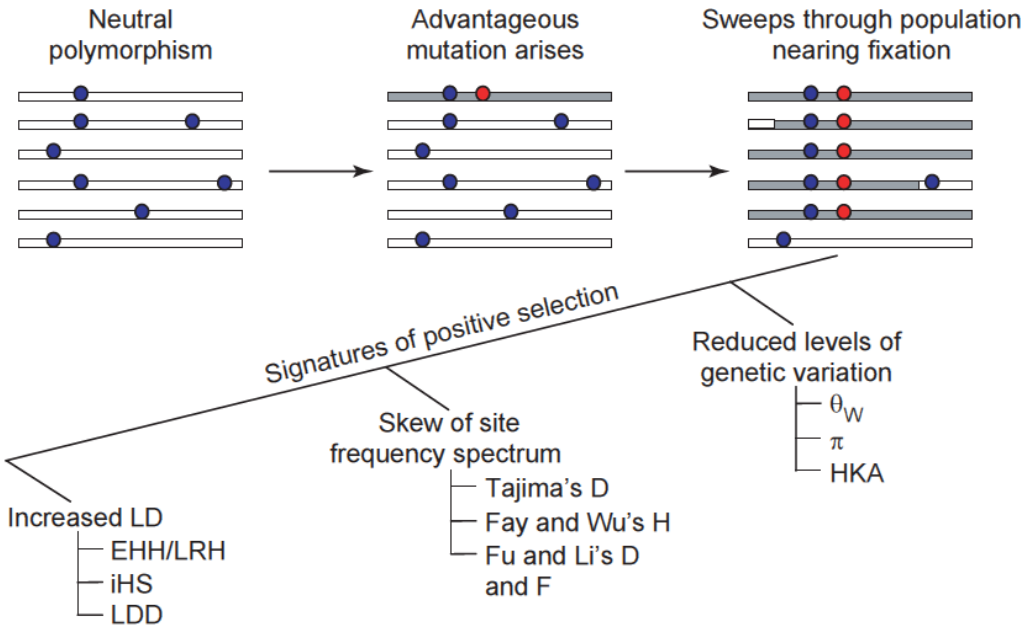
Genotyping of medically important variants is not only useful for predicting the likelihood of adverse reactions as a result of treatment with specific drugs, but also in predicting disease risks. For *CYP3A5* variability, inter-ethnic differences in the risk of developing salt-sensitive hypertension are well established (Givens et al. 2003; Brown 2006; Bochud et al. 2009). Adequate salt and water retention is beneficial in populations which are found close to the equator. The high frequencies of *CYP3A5*3* in Ethiopia may be causing conditions such as hyponatremia (abnormally low levels of sodium in the blood which is associated with dehydration). The Ethiopian data re-sequenced as part of this study suggest that there may be differential susceptibilities to hyponatremia within the country. An interesting study could be to examine differences in the prevalence of hyponatremia between the five Ethiopian groups and examine whether the condition is a) frequent within Ethiopia, b) observed at different frequencies in Ethiopia and c) associated with the *CYP3A5*3* allele. An

additional study which would be interesting is to examine whether there are misdiagnoses of prevalent infectious and tropical diseases, such as malaria, which are actually hyponatremia. Common phenotypic effects of hyponatremia include nausea, vomiting, headaches, confusion, lethargy, fatigue and appetite loss; all symptoms; which are also associated with malaria infection (Cook et al. 2003). Malaria is highly prevalent within Ethiopia (Karunamoorthi and Bekele 2009). During diagnosis of infection, if a patient presents with malaria-like symptoms, but has a negative smear for parasites, they are often treated with anti-malarials. It would be interesting to see whether a large number of these cases are actually hyponatremia.

8.4. There is strong evidence of positive selection for low/non CYP3A5 expression

The "genomics" revolution will undoubtedly aid studies which look for genomic targets of positive selection. Whole genome scans and studies of individual genes can identify genomic regions which are likely to have undergone positive selection. Signatures of positive selection are often seen following a selective sweep and can be detected by a number of different statistics (see Figure 8.1).

Figure 8.1: Signatures of positive selection. Neutral polymorphisms (shown in blue) are driven to high frequency as a result of being in high LD with a selected, advantageous allele (shown in red). The resulting selective sweep leads to a reduction in polymorphism levels in haplotypes surrounding the selected region. Reduction in polymorphism levels are detected by statistics which include nucleotide polymorphism levels (θ_w), nucleotide diversity (π) and the HKA test. Additional statistics detect an excess of rare variants (Tajima's D and Fu and Li's D and F), high frequency derived variants (Fay and Wu's H). Finally, extended regions of linkage disequilibrium, relative to neutral expectations are also seen following a selective sweep. The image has been taken from (Biswas and Akey 2006).



The data generated for this thesis complement the results of previous publications which reported evidence of geographically restricted positive selection on the *CYP3A5* gene; specifically on the *CYP3A5*3* low/non-expresser allele (Thompson et al. 2004; Biswas and Akey 2006; Li et al. 2011). Few signatures of positive selection were observed in the African populations re-sequenced for this thesis, outside of East Africa. Of course examination of the entire *CYP3A5* region in different African groups will enable accurate comparisons with non-African populations regarding selection.

Prior to work presented in this thesis the age of the *CYP3A5*3* allele had not been estimated. This thesis estimated that the *CYP3A5*3* mutation first arose ~76,000 years ago (chapter 6). This date is consistent with its presence in and outside of Africa. These data provide strong evidence of differential selection on the *CYP3A5*3* allele. An examination of the different frequencies within and outside of Africa, and the significant positive correlation between latitude and *CYP3A5*3* allele frequencies, strongly support a hypothesis that the allele is advantageous outside of the African continent. The paucity of variation observed on *CYP3A5*3* haplotype backgrounds in non-African populations suggest that the allele swept to high frequency rapidly outside of Africa; perhaps soon after it first arose. The age of the allele is consistent with the estimates of when modern humans first appeared in West Africa and in the Arabian Peninsula (see chapter 7). The stark contrast in *CYP3A5*3* frequencies between these two geographic regions suggest that some differential environmental pressures have selected for *CYP3A5*3* and driven it to high frequency. Of course, single gene approaches to detecting evidence for positive selection do not have the same power as scans of large genomic regions where the extent of LD decay can be visualised. It is possible that the *CYP3A5*3* allele is one of multiple targets of selection on chromosome 7 and has hitchhiked to high frequency on the back of advantageous mutations (Hofer et al. 2009).

The data presented in this thesis highlight the importance of appreciating population history and of utilising evolutionary approaches in clinical research in order to elucidate mechanisms which a) cause differentiation of patient populations and b) identify patient populations which are likely to be distinct from a wider cohort because of differences in human evolutionary history. For example it has been well established that hypertension incidence and prevalence is highest in populations with recent African ancestry (Cooper et al. 1997; Rotimi and Jorde 2010). Evolutionary approaches to understanding hypertension have identified that inter-ethnic differences in the risk of developing the disease are linked to important events in human evolutionary, and migratory, history (Young et al. 2005). This in turn can aid in identifying groups which are at risk of elevated hypertension. Often medical studies have focused on the clinical implications of particular genes and overlooked

evolutionary aspects which can lead to population differentiation. Evolutionary based approaches to genes, such as *CYP3A5*, will identify populations where alternative drug regimens, to those in current widespread use, are likely to be required. Stratified, or personalised, medicine is a key goal of the genomics revolution (Samani et al. 2010) and evolutionary factors are likely to be essential in identifying patients who require idiosyncratic interventions.

9. Future work

All lab work and analyses presented in this thesis have been performed within the time frame for completing a Ph.D. thesis and using the funds and resources available. However the work also highlighted a number of areas for additional research, to understand intra-African diversity at the *CYP3A5* locus and to examine the work in the context of medical and evolutionary research. The main areas for future research are discussed below.

9.1. African re-sequencing

This thesis could not elucidate intra-African diversity for the entire *CYP3A5* gene, due to funding constraints with re-sequencing of non-Ethiopian groups. A comparison of the full gene sequence may identify more African populations which have *CYP3A5* haplotype structures similar to those seen in Ethiopia or other non-African populations. This would enable a comprehensive comparison of intra-African diversity over the entire *CYP3A5* gene. Additionally re-sequencing of a larger *CYP3A5* region may identify novel variants, at appreciable frequencies, likely to affect CYP3A5 expression, and thus disease susceptibility and risk of adverse drug reactions, in Africa.

9.2. Assessing the functional implications of CYP3A5 variants

Bioinformatics analysis in chapter 5 identified a number of variants which may potentially affect CYP3A5 protein expression levels within Africa. Bioinformatic predictions alone do not provide tangible evidence of what occurs *in vivo*, unlike *in vitro* splicing assays (discussed in 5.4.1 and 5.4.2). Of all identified variants, a novel intronic 10bp deletion, identified in five individuals from West Central Africa, is an immediate candidate for functional studies. Although bioinformatics analysis did not predict that this polymorphism would affect intron 1 splicing, its proximity to the intron acceptor splice site makes it an ideal candidate for functional analysis.

9.3. Further population comparisons and simulations

There has been a rapid increase in the amount of genomic data available online. The 1000 Genomes data are likely to provide one of the most comprehensive datasets of human genetic variation. High coverage data from the project will be invaluable for assessing global population differences from re-sequencing data. A comparison of the African data generated for this thesis with soon-to-be released high coverage data from the 1000 Genomes Project

will enable analysis of African diversity in a wider global context than the current data available online.

Demographic and environmental factors such as migration, population admixture and latitude could be modelled within simulated datasets (Ritchie and Bush 2010). Comparisons of simulated datasets which closely represent the observed data can infer the likely demographic and environmental factors which have lead to particular genomic patterns. Simulations of population histories could be performed to examine African *CYP3A5* diversity in greater detail. This would also help to elucidate how *CYP3A5* haplotype patterns observed in Ethiopia are likely to have arisen.

9.4. Examining medical associations of CYP3A5 variability

As outlined in the Introduction to this thesis variability in CYP3A5 expression is associated with differential susceptibility to disease and adverse clinical outcomes. This study has characterised variation within the entire *CYP3A5* gene. A study of variation in a medically important gene is useful when it is also considered for its implications on clinical outcomes. There is potential to perform case-control studies comparing clinical outcomes, associated with specific drug substrates, in different populations to see how the identified variants may affect clinical outcomes. This thesis has highlighted differences between East Africans and other sub-Saharan African groups which are likely to affect treatment outcomes. Case-control studies comparing inter-population differences, for example between North-Eastern Ethiopian groups, such as the Amhara, with the Anuak and Bantu-speaking populations, separately, in their ability to metabolise well-characterised CYP3A substrates such as midazolam; and drugs used to treat HIV-1 and malaria, as well those used in mass drug administration campaigns (one of the major ways in which African populations have access to medical treatment) would identify individuals at risk of adverse clinical outcomes within specific populations. Additional studies of the association between *CYP3A5*3* and hyponatremia in equatorial Ethiopia would also be of medical benefit (see section 8.3).

9.5. Re-sequencing of other CYP450 genes

There are multiple *CYP450* genes which are responsible for the metabolism of a wide range of drugs in clinical use. Re-sequencing of further *CYP450* genes, and others involved in drug metabolism could potentially characterise medically important variation. Additionally, characterising variation in genes which are also involved in the metabolism of many CYP3A5 substrates may also provide further information on how multiple drug metabolising enzymes

interact and how variation in one gene, encoding an important enzyme, can affect a drug metabolising pathway.

9.6. Re-sequencing of a larger region of chromosome 7 surrounding the CYP3A5 gene

One of the very interesting areas of research that follow on from this thesis is to examine a larger region of chromosome 7 within global populations, and also in Africa, to elucidate whether *CYP3A5* is likely to have been a target of selection within the region. Within Africa this would be useful for haplotypes defined by the *CYP3A5*6* variant. The LRH test is designed to detect recent selective sweeps (within the past ~30,000 years). Therefore, from data presented in this thesis, it cannot be ruled out that this allele had been the target of selection prior to this time frame. Additionally, the design of composite likelihood models to examine statistically significant similarities and differences in iHS and XP-EHH estimates would provide better indications of how likely selection signatures are indicative of true selection events.

9.7. Concluding remarks

This thesis has characterised novel and known variation in a medically important gene which could have important clinical implications for healthcare intervention strategies. The African data provide an example of the importance of including diverse populations in genomics, medical and evolutionary research. Additionally, the work presented on *CYP3A5* has come at a time where the genomics revolution is generating huge amounts of data on multiple global populations; including those in Africa. This will inevitably improve the design of studies which look for medically important variation, and may improve the provision of healthcare for patients with recent African ancestry. The UK and EU are slowly integrating personalised medicine into standard healthcare practices (Coggi 2011; Fricker 2011). It is possible that studies, such as those conducted for this thesis, will assist in genotype or genome guided medicine; either for the individual or population.

10. References

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev (2010). "A method and server for predicting damaging missense mutations." Nature methods **7**(4): 248-249.

Aitman, T. J., L. D. Cooper, P. J. Norsworthy, F. N. Wahid, J. K. Gray, B. R. Curtis, P. M. McKeigue, D. Kwiatkowski, B. M. Greenwood, R. W. Snow, A. V. Hill and J. Scott (2000). "Malaria susceptibility and CD36 mutation." Nature **405**(6790): 1015-1016.

Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.

Alt, F. W., A. L. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland and D. Baltimore (1980). "Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends." Cell **20**(2): 293-301.

Altshuler, D. M., R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, P. E. Bonnen, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, C. Gonzaga-Jauregui, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, Q. Zhang, M. J. Ghori, R. McGinnis, W. McLaren, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks and J. E. McEwen (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.

Anttila, S., J. Hukkanen, J. Hakkola, T. Stjernvall, P. Beaune, R. J. Edwards, A. R. Boobis, O. Pelkonen and H. Raunio (1997). "Expression and localization of CYP3A4 and CYP3A5 in human lung." Am J Respir Cell Mol Biol **16**(3): 242-249.

Aoyama, T., S. Yamano, D. J. Waxman, D. P. Lapenson, U. A. Meyer, V. Fischer, R. Tyndale, T. Inaba, W. Kalow, H. V. Gelboin and et al. (1989). "Cytochrome P-450 hPCN3, a novel cytochrome P-450 IIIA gene product that is differentially expressed in adult human liver. cDNA and deduced amino acid sequence and distinct specificities of cDNA-expressed hPCN1 and hPCN3 for the metabolism of steroid hormones and cyclosporine." J Biol Chem **264**(18): 10388-10395.

Aspray, T. J., H. Kitange, P. Setel, N. C. Unwin and D. Whiting (1998). "Disease burden in sub-Saharan Africa." Lancet **351**(9110): 1208-1209.

Bachtrog, D. (2008). "Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes." BMC evolutionary biology **8**: 334.

Balding, D. J. (2006). "A tutorial on statistical methods for population association studies." Nat Rev Genet **7**(10): 781-791.

Balram, C., Q. Zhou, Y. B. Cheung and E. J. Lee (2003). "CYP3A5*3 and *6 single nucleotide polymorphisms in three distinct Asian populations." Eur J Clin Pharmacol **59**(2): 123-126.

Bandelt, H. J., P. Forster and A. Rohl (1999). "Median-joining networks for inferring intraspecific phylogenies." Molecular biology and evolution **16**(1): 37-48.

Bandelt, H. J., P. Forster, B. C. Sykes and M. B. Richards (1995). "Mitochondrial portraits of human populations using median networks." Genetics **141**(2): 743-753.

Barbujani, G., A. Russo, A. Farabegoli and E. Calzolari (1989). "Inferences on the inheritance of congenital anomalies from temporal and spatial patterns of occurrence." Genetic epidemiology **6**(4): 537-552.

Behar, D. M., M. G. Thomas, K. Skorecki, M. F. Hammer, E. Bulygina, D. Rosengarten, A. L. Jones, K. Held, V. Moses, D. Goldstein, N. Bradman and M. E. Weale (2003). "Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries." American journal of human genetics **73**(4): 768-779.

Beleza, S., L. Gusmao, A. Amorim, A. Carracedo and A. Salas (2005). "The genetic legacy of western Bantu migrations." Hum Genet **117**(4): 366-375.

Bennett, S. T., C. Barnes, A. Cox, L. Davies and C. Brown (2005). "Toward the 1,000 dollars human genome." Pharmacogenomics **6**(4): 373-382.

Berniell-Lee, G., F. Calafell, E. Bosch, E. Heyer, L. Sica, P. Mouguiama-Daouda, L. van der Veen, J. M. Hombert, L. Quintana-Murci and D. Comas (2009). "Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages." Mol Biol Evol **26**(7): 1581-1589.

Biswas, S. and J. M. Akey (2006). "Genomic insights into positive selection." Trends in genetics : TIG **22**(8): 437-446.

Bochud, M., P. Bovet, M. Burnier and C. B. Eap (2009). "CYP3A5 and ABCB1 genes and hypertension." Pharmacogenomics **10**(3): 477-487.

Bochud, M., C. B. Eap, R. C. Elston, P. Bovet, M. Maillard, L. Schild, C. Shamlaye and M. Burnier (2006). "Association of CYP3A5 genotypes with blood pressure and renal function in African families." J Hypertens **24**(5): 923-929.

Boobis, A. R., R. J. Edwards, D. A. Adams and D. S. Davies (1996). "Dissecting the function of cytochrome P450." Br J Clin Pharmacol **42**(1): 81-89.

Brockmoller, J., J. Kirchheiner, C. Meisel and I. Roots (2000). "Pharmacogenetic diagnostics of cytochrome P450 polymorphisms in clinical drug development and in drug treatment." Pharmacogenomics **1**(2): 125-151.

Brown, M. J. (2006). "Hypertension and ethnic group." BMJ **332**(7545): 833-836.

Browning, S. L., A. Tarekegn, E. Bekele, N. Bradman and M. G. Thomas (2010). "CYP1A2 is more variable than previously thought: a genomic biography of the gene behind the human drug-metabolizing enzyme." Pharmacogenetics and genomics **20**(11): 647-664.

Burk, O., I. Koch, J. Raucy, E. Hustert, M. Eichelbaum, J. Brockmoller, U. M. Zanger and L. Wojnowski (2004). "The induction of cytochrome P450 3A5 (CYP3A5) in the human liver and intestine is mediated by the xenobiotic sensors pregnane X receptor (PXR) and constitutively activated receptor (CAR)." The Journal of biological chemistry **279**(37): 38379-38385.

Busi, F. and T. Cresteil (2005). "CYP3A5 mRNA degradation by nonsense-mediated mRNA decay." Mol Pharmacol **68**(3): 808-815.

Butler, D. (2001). "Genomics. Are you ready for the revolution?" Nature **409**(6822): 758-760.

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

Campbell, M. C. and S. A. Tishkoff (2010). "The evolution of human genetic and phenotypic variation in Africa." Current biology : CB **20**(4): R166-173.

Cartharius, K., K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein and T. Werner (2005). "MatInspector and beyond: promoter analysis based on transcription factor binding sites." Bioinformatics **21**(13): 2933-2942.

Cattaneo, D., S. Baldelli and N. Perico (2008). "Pharmacogenetics of immunosuppressants: progress, pitfalls and promises." Am J Transplant **8**(7): 1374-1383.

Chatterjee, S. and J. K. Pal (2009). "Role of 5'- and 3'-untranslated regions of mRNAs in human diseases." Biology of the cell / under the auspices of the European Cell Biology Organization **101**(5): 251-262.

Chen, X., H. Wang, G. Zhou, X. Zhang, X. Dong, L. Zhi, L. Jin and F. He (2009). "Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine." Environ Health Perspect **117**(10): 1541-1548.

Chou, F. C., S. J. Tzeng and J. D. Huang (2001). "Genetic polymorphism of cytochrome P450 3A5 in Chinese." Drug Metab Dispos **29**(9): 1205-1209.

Chou, F. C., S. J. Tzeng and J. D. Huang (2001). "Genetic polymorphism of cytochrome P450 3A5 in Chinese." Drug metabolism and disposition: the biological fate of chemicals **29**(9): 1205-1209.

Chung, R. T. (2010). "Reaping the early harvest of the genomics revolution." Gastroenterology **138**(5): 1653-1654.

Coggi, P. T. (2011). "A European view on the future of personalised medicine in the EU." European journal of public health **21**(1): 6-7.

Coleman, R. (1998). "Disease burden in sub-Saharan Africa." Lancet **351**(9110): 1208.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, N. A. Rosenberg and J. K. Pritchard (2006). "A worldwide survey of haplotype variation and linkage disequilibrium in the human genome." Nat Genet **38**(11): 1251-1260.

Constable, S., M. R. Johnson and M. Pirmohamed (2006). "Pharmacogenetics in clinical practice: considerations for testing." Expert Rev Mol Diagn **6**(2): 193-205.

Cook, G. C., A. Zumla and P. Manson (2003). Manson's tropical diseases. Edinburgh, Saunders.

Cook, G. C., A. Zumla and P. Manson (2009). Manson's tropical diseases. Edinburgh, Saunders/Elsevier.

Coop, G., D. Witonsky, A. Di Rienzo and J. K. Pritchard (2010). "Using environmental correlations to identify loci underlying local adaptation." Genetics **185**(4): 1411-1423.

Cooper, R., C. Rotimi, S. Ataman, D. McGee, B. Osotimehin, S. Kadiri, W. Muna, S. Kingue, H. Fraser, T. Forrester, F. Bennett and R. Wilks (1997). "The prevalence of hypertension in seven populations of west African origin." American journal of public health **87**(2): 160-168.

Cruciani, F., P. Santolamazza, P. Shen, V. Macaulay, P. Moral, A. Olckers, D. Modiano, S. Holmes, G. Destro-Bisol, V. Coia, D. C. Wallace, P. J. Oefner, A. Torroni, L. L. Cavalli-Sforza, R. Scozzari and P. A. Underhill (2002). "A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes." American journal of human genetics **70**(5): 1197-1214.

Cserti-Gazdewich, C. M., W. R. Mayr and W. H. Dzik (2011). "Plasmodium falciparum malaria and the immunogenetics of ABO, HLA, and CD36 (platelet glycoprotein IV)." Vox sanguinis **100**(1): 99-111.

Dally, H., H. Bartsch, B. Jager, L. Edler, P. Schmezer, B. Spiegelhalder, H. Dienemann, P. Drings, K. Kayser, V. Schulz and A. Risch (2004). "Genotype relationships in the CYP3A locus in Caucasians." Cancer Lett **207**(1): 95-99.

Dandara, C., R. Ballo and M. I. Parker (2005). "CYP3A5 genotypes and risk of oesophageal cancer in two South African populations." Cancer Lett **225**(2): 275-282.

de Filippo, C., C. Barbieri, M. Whitten, S. W. Mpoloka, E. D. Gunnarsdottir, K. Bostoen, T. Nyambe, K. Beyer, H. Schreiber, P. de Knijff, D. Luiselli, M. Stoneking and B. Pakendorf "Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups." Mol Biol Evol **28**(3): 1255-1269.

De Stefano, G. F., C. Martinez-Labarga, R. Casalotti, M. Tartaglia, A. Novelletto, G. Pepe and O. Rickards (2002). "Analysis of three RFLPs of the COL1A2 (Type I Collagen) in the Amhara and the Oromo of Ethiopia." Ann Hum Biol **29**(4): 432-441.

Defrancesco, L. (2012). "Life Technologies promises $1,000 genome." Nature biotechnology **30**(2): 126.

DeGiorgio, M., M. Jakobsson and N. A. Rosenberg (2009). "Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa." Proc Natl Acad Sci U S A **106**(38): 16057-16062.

Destro-Bisol, G., R. Maviglia, A. Caglia, I. Boschi, G. Spedini, V. Pascali, A. Clark and S. Tishkoff (1999). "Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu)." Hum Genet **104**(2): 149-157.

Di Rienzo, A. and R. R. Hudson (2005). "An evolutionary framework for common diseases: the ancestral-susceptibility model." Trends in genetics : TIG **21**(11): 596-601.

Diczfalusy, U., J. Miura, H. K. Roh, R. A. Mirghani, J. Sayi, H. Larsson, K. G. Bodin, A. Allqvist, M. Jande, J. W. Kim, E. Aklillu, L. L. Gustafsson and L. Bertilsson (2008). "4Beta-hydroxycholesterol is a new endogenous CYP3A marker: relationship to CYP3A5 genotype, quinine 3-hydroxylation and sex in Koreans, Swedes and Tanzanians." Pharmacogenet Genomics **18**(3): 201-208.

Eap, C. B., T. Buclin, E. Hustert, G. Bleiber, K. P. Golay, A. C. Aubert, P. Baumann, A. Telenti and R. Kerb (2004). "Pharmacokinetics of midazolam in CYP3A4- and CYP3A5-genotyped subjects." Eur J Clin Pharmacol **60**(4): 231-236.

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

Excoffier, L. and M. Slatkin (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." Molecular biology and evolution **12**(5): 921-927.

Farrall, M. and D. E. Weeks (1998). "Mutational mechanisms for generating microsatellite allele-frequency distributions: an analysis of 4,558 markers." American journal of human genetics **62**(5): 1260-1262.

Fay, J. C. and C. I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics **155**(3): 1405-1413.

Fellay, J., C. Marzolini, L. Decosterd, K. P. Golay, P. Baumann, T. Buclin, A. Telenti and C. B. Eap (2005). "Variations of CYP3A activity induced by antiretroviral treatment in HIV-1 infected patients." Eur J Clin Pharmacol **60**(12): 865-873.

Ferreira, P. E., M. I. Veiga, I. Cavaco, J. P. Martins, B. Andersson, S. Mushin, A. S. Ali, A. Bhattarai, V. Ribeiro, A. Bjorkman and J. P. Gil (2008). "Polymorphism of antimalaria drug metabolizing, nuclear receptor, and drug transport genes among malaria patients in Zanzibar, East Africa." Ther Drug Monit **30**(1): 10-15.

Finta, C. and P. G. Zaphiropoulos (2000). "The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons." Gene **260**(1-2): 13-23.

Floyd, M. D., G. Gervasini, A. L. Masica, G. Mayo, A. L. George, Jr., K. Bhat, R. B. Kim and G. R. Wilkinson (2003). "Genotype-phenotype associations for common CYP3A4 and CYP3A5 variants in the basal and induced metabolism of midazolam in European- and African-American men and women." Pharmacogenetics **13**(10): 595-606.

Foti, R. S. and M. B. Fisher (2004). "Importance of patient selection when determining the significance of the CYP3A5 polymorphism in clinical trials." Pharmacogenomics J **4**(6): 362-364.

Freeman, B., N. Smith, C. Curtis, L. Huckett, J. Mill and I. W. Craig (2003). "DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping." Behavior genetics **33**(1): 67-72.

Fricker, J. (2011). "UK's adopts systematic approach to personalised cancer medicine." Molecular oncology **5**(3): 217-219.

Friedrich, C. A. (2001). "Genomics 101: what the practicing physician needs to know about the genetics revolution. June 8-9, 2001, Chicago, IL, USA." Expert review of molecular diagnostics **1**(2): 135-136.

Frisch, A., R. Colombo, E. Michaelovsky, M. Karpati, B. Goldman and L. Peleg (2004). "Origin and spread of the 1278insTATC mutation causing Tay-Sachs disease in Ashkenazi Jews: genetic drift as a robust and parsimonious hypothesis." Hum Genet **114**(4): 366-376.

Frohlich, M., M. M. Hoffmann, J. Burhenne, G. Mikus, J. Weiss and W. E. Haefeli (2004). "Association of the CYP3A5 A6986G (CYP3A5*3) polymorphism with saquinavir pharmacokinetics." Br J Clin Pharmacol **58**(4): 443-444.

Fromm, M. F., B. M. Schmidt, A. Pahl, J. Jacobi and R. E. Schmieder (2005). "CYP3A5 genotype is associated with elevated blood pressure." Pharmacogenet Genomics **15**(10): 737-741.

Fromm, M. F., H. Schwilden, I. Bachmakov, J. Konig, F. Bremer and J. Schuttler (2007). "Impact of the CYP3A5 genotype on midazolam pharmacokinetics and pharmacodynamics during intensive care sedation." Eur J Clin Pharmacol **63**(12): 1129-1133.

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

Fu, Y. X. and W. H. Li (1993). "Statistical tests of neutrality of mutations." Genetics **133**(3): 693-709.

Fukuen, S., T. Fukuda, H. Maune, Y. Ikenaga, I. Yamamoto, T. Inaba and J. Azuma (2002). "Novel detection assay by PCR-RFLP and frequency of the CYP3A5 SNPs, CYP3A5*3 and *6, in a Japanese population." Pharmacogenetics **12**(4): 331-334.

Gamazon, E. R., W. Zhang, R. S. Huang, M. E. Dolan and N. J. Cox (2009). "A pharmacogene database enhanced by the 1000 Genomes Project." Pharmacogenetics and genomics **19**(10): 829-832.

Gardiner, S. J. and E. J. Begg (2006). "Pharmacogenetics, drug-metabolizing enzymes, and clinical practice." Pharmacological reviews **58**(3): 521-590.

Gebeyehu, E., E. Engidawork, A. Bijnsdorp, A. Aminy, U. Diczfalusy and E. Aklillu "Sex and CYP3A5 genotype influence total CYP3A activity: high CYP3A activity and a unique distribution of CYP3A5 variant alleles in Ethiopians." Pharmacogenomics J.

Gebeyehu, E., E. Engidawork, A. Bijnsdorp, A. Aminy, U. Diczfalusy and E. Aklillu (2011). "Sex and CYP3A5 genotype influence total CYP3A activity: high CYP3A activity and a unique distribution of CYP3A5 variant alleles in Ethiopians." The pharmacogenomics journal **11**(2): 130-137.

Gerbault, P., A. Liebert, Y. Itan, A. Powell, M. Currat, J. Burger, D. M. Swallow and M. G. Thomas (2011). "Evolution of lactase persistence: an example of human niche construction." Philosophical transactions of the Royal Society of London. Series B, Biological sciences **366**(1566): 863-877.

Getachew, K. N. (1998). "Tradition, continuity and socio-economic change among the pastoral Afar of the Middle Awash Valley in North Eastern Ethiopia. [electronic resource]."

Getachew, K. N. (2002). Among the pastoral afar in Ethiopia : tradition, continuity and socio-economic change. Utrecht, International Books ; Charlbury : Jon Carpenter.

Ghosh, D. (2000). "Object-oriented transcription factors database (ooTFD)." Nucleic acids research **28**(1): 308-310.

Ghosh, S. S., A. K. Basu, S. Ghosh, R. Hagley, L. Kramer, J. Schuetz, W. M. Grogan, P. Guzelian and C. O. Watlington (1995). "Renal and hepatic family 3A cytochromes P450 (CYP3A) in spontaneously hypertensive rats." Biochemical pharmacology **50**(1): 49-54.

Giacomini, K. M., C. M. Brett, R. B. Altman, N. L. Benowitz, M. E. Dolan, D. A. Flockhart, J. A. Johnson, D. F. Hayes, T. Klein, R. M. Krauss, D. L. Kroetz, H. L. McLeod, A. T. Nguyen, M. J. Ratain, M. V. Relling, V. Reus, D. M. Roden, C. A. Schaefer, A. R. Shuldiner, T. Skaar, K. Tantisira, R. F. Tyndale, L. Wang, R. M. Weinshilboum, S. T. Weiss and I. Zineh (2007). "The pharmacogenetics research network: from SNP discovery to clinical drug response." Clin Pharmacol Ther **81**(3): 328-345.

Givens, R. C., Y. S. Lin, A. L. Dowling, K. E. Thummel, J. K. Lamba, E. G. Schuetz, P. W. Stewart and P. B. Watkins (2003). "CYP3A5 genotype predicts renal CYP3A activity and blood pressure in healthy adults." J Appl Physiol **95**(3): 1297-1300.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman (1995). "Genetic absolute dating based on microsatellites and the origin of modern humans." Proceedings of the National Academy of Sciences of the United States of America **92**(15): 6723-6727.

Goldstein, D. B. and M. E. Weale (2001). "Population genomics: linkage disequilibrium holds the key." Curr Biol **11**(14): R576-579.

Gonzalez, F. J. and H. V. Gelboin (1994). "Role of human cytochromes P450 in the metabolic activation of chemical carcinogens and toxins." Drug Metab Rev **26**(1-2): 165-183.

Goode, D. L., G. M. Cooper, J. Schmutz, M. Dickson, E. Gonzales, M. Tsai, K. Karra, E. Davydov, S. Batzoglou, R. M. Myers and A. Sidow (2010). "Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes." Genome research **20**(3): 301-310.

Goto, M., S. Masuda, T. Kiuchi, Y. Ogura, F. Oike, M. Okuda, K. Tanaka and K. Inui (2004). "CYP3A5*1-carrying graft liver reduces the concentration/oral dose ratio of tacrolimus in recipients of living-donor liver transplantation." Pharmacogenetics **14**(7): 471-478.

Graur, D., W.-H. Li and W.-H. F. o. m. e. Li (2000). Fundamentals of molecular evolution. Sunderland, Mass., Sinauer Associates.

Griffiths, R. C. and P. Marjoram (1996). "Ancestral inference from samples of DNA sequences with recombination." Journal of computational biology : a journal of computational molecular cell biology **3**(4): 479-502.

Guo, S. W. and E. A. Thompson (1992). "Performing the exact test of Hardy-Weinberg proportion for multiple alleles." Biometrics **48**(2): 361-372.

Haas, D. W., J. A. Bartlett, J. W. Andersen, I. Sanne, G. R. Wilkinson, J. Hinkle, F. Rousseau, C. D. Ingram, A. Shaw, M. M. Lederman and R. B. Kim (2006). "Pharmacogenetics of nevirapine-associated hepatotoxicity: an Adult AIDS Clinical Trials Group collaboration." Clin Infect Dis **43**(6): 783-786.

Hammer, M. F., A. J. Redd, E. T. Wood, M. R. Bonner, H. Jarjanazi, T. Karafet, S. Santachiara-Benerecetti, A. Oppenheim, M. A. Jobling, T. Jenkins, H. Ostrer and B. Bonne-Tamir (2000). "Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes." Proc Natl Acad Sci U S A **97**(12): 6769-6774.

Hedrick, P. W. (2007). "Sex: differences in mutation, recombination, selection, gene flow, and genetic drift." Evolution; international journal of organic evolution **61**(12): 2750-2771.

Henze, P. B. (2000). Layers of time : a history of Ethiopia. New York, St. Martin's Press.

Hesselink, D. A., R. H. van Schaik, I. P. van der Heiden, M. van der Werf, P. J. Gregoor, J. Lindemans, W. Weimar and T. van Gelder (2003). "Genetic polymorphisms of the CYP3A4, CYP3A5, and MDR-1 genes and pharmacokinetics of the calcineurin inhibitors cyclosporine and tacrolimus." Clin Pharmacol Ther **74**(3): 245-254.

Hofer, T., N. Ray, D. Wegmann and L. Excoffier (2009). "Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection." Annals of human genetics **73**(1): 95-108.

Hollox, E. J., M. Poulter, M. Zvarik, V. Ferak, A. Krause, T. Jenkins, N. Saha, A. I. Kozlov and D. M. Swallow (2001). "Lactase haplotype diversity in the Old World." Am J Hum Genet **68**(1): 160-172.

Horsfall, L. J., D. Zeitlyn, A. Tarekegn, E. Bekele, M. G. Thomas, N. Bradman and D. M. Swallow (2011). "Prevalence of clinically relevant UGT1A alleles and haplotypes in African populations." Annals of human genetics **75**(2): 236-246.

Hotez, P. J. (2009). "Mass drug administration and integrated control for the world's high-prevalence neglected tropical diseases." Clin Pharmacol Ther **85**(6): 659-664.

Hotez, P. J. and A. Kamath (2009). "Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden." PLoS neglected tropical diseases **3**(8): e412.

Huff, C. D., H. C. Harpending and A. R. Rogers (2010). "Detecting positive selection from genome scans of linkage disequilibrium." BMC Genomics **11**: 8.

Hughes, T. A. (2006). "Regulation of gene expression by alternative untranslated regions." Trends in genetics : TIG **22**(3): 119-122.

Hukkanen, J., T. Vaisanen, A. Lassila, R. Piipari, S. Anttila, O. Pelkonen, H. Raunio and J. Hakkola (2003). "Regulation of CYP3A5 by glucocorticoids and cigarette smoke in human lung-derived cells." J Pharmacol Exp Ther **304**(2): 745-752.

Hustert, E., M. Haberl, O. Burk, R. Wolbold, Y. Q. He, K. Klein, A. C. Nuessler, P. Neuhaus, J. Klattig, R. Eiselt, I. Koch, A. Zibat, J. Brockmoller, J. R. Halpert, U. M. Zanger and L. Wojnowski (2001). "The genetic determinants of the CYP3A5 polymorphism." Pharmacogenetics **11**(9): 773-779.

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

Ingram, C. J., M. F. Elamin, C. A. Mulcare, M. E. Weale, A. Tarekegn, T. O. Raga, E. Bekele, F. M. Elamin, M. G. Thomas, N. Bradman and D. M. Swallow (2007). "A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence?" Hum Genet **120**(6): 779-788.

Ingram, C. J., C. A. Mulcare, Y. Itan, M. G. Thomas and D. M. Swallow (2009). "Lactose digestion and the evolutionary genetics of lactase persistence." Hum Genet **124**(6): 579-591.

Ingram, C. J., T. O. Raga, A. Tarekegn, S. L. Browning, M. F. Elamin, E. Bekele, M. G. Thomas, M. E. Weale, N. Bradman and D. M. Swallow (2009). "Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group." J Mol Evol **69**(6): 579-588.

Iwano, S., T. Saito, Y. Takahashi, K. Fujita and T. Kamataki (2001). "Cooperative regulation of CYP3A5 gene transcription by NF-Y and Sp family members." Biochemical and biophysical research communications **286**(1): 55-60.

Iwasaki, K. (2007). "Metabolism of tacrolimus (FK506) and recent topics in clinical pharmacokinetics." Drug Metab Pharmacokinet **22**(5): 328-335.

Jobling, M. A., M. Hurles and C. Tyler-Smith (2004). Human evolutionary genetics : origins, peoples & disease. New York, NY ; London, Garland.

Johnson, J. A. (2008). "Ethnic differences in cardiovascular drug response: potential contribution of pharmacogenetics." Circulation **118**(13): 1383-1393.

Jones, B. L. and D. M. Swallow (2011). "The impact of cis-acting polymorphisms on the human phenotype." The HUGO Journal.

Josephson, F., A. Allqvist, M. Janabi, J. Sayi, E. Aklillu, M. Jande, M. Mahindi, J. Burhenne, Y. Bottiger, L. L. Gustafsson, W. E. Haefeli and L. Bertilsson (2007). "CYP3A5 genotype has an impact on the metabolism of the HIV protease inhibitor saquinavir." Clin Pharmacol Ther **81**(5): 708-712.

Jounaidi, Y., P. S. Guzelian, P. Maurel and M. J. Vilarem (1994). "Sequence of the 5'-flanking region of CYP3A5: comparative analysis with CYP3A4 and CYP3A7." Biochemical and biophysical research communications **205**(3): 1741-1747.

Jounaidi, Y., V. Hyrailles, L. Gervot and P. Maurel (1996). "Detection of CYP3A5 allelic variant: a candidate for the polymorphic expression of the protein?" Biochemical and biophysical research communications **221**(2): 466-470.

Kaiser, J. (2008). "DNA sequencing. A plan to capture human diversity in 1000 genomes." Science **319**(5862): 395.

Kang, R. H., S. M. Jung, K. A. Kim, D. K. Lee, H. K. Cho, B. J. Jung, Y. K. Kim, S. H. Kim, C. Han, M. S. Lee and J. Y. Park (2009). "Effects of CYP2D6 and CYP3A5 genotypes on the plasma concentrations of

risperidone and 9-hydroxyrisperidone in Korean schizophrenic patients." J Clin Psychopharmacol **29**(3): 272-277.

Karunamoorthi, K. and M. Bekele (2009). "Prevalence of malaria from peripheral blood smears examination: a 1-year retrospective study from the Serbo Health Center, Kersa Woreda, Ethiopia." Journal of infection and public health **2**(4): 171-176.

Kelley, J. L. and W. J. Swanson (2008). "Positive selection in the human genome: from genome scans to biological significance." Annual review of genomics and human genetics **9**: 143-160.

Kim, K. A., P. W. Park, K. H. Liu, K. B. Kim, H. J. Lee, J. G. Shin and J. Y. Park (2008). "Effect of rifampin, an inducer of CYP3A and P-glycoprotein, on the pharmacokinetics of risperidone." Journal of clinical pharmacology **48**(1): 66-72.

Kim, R. B., C. Wandel, B. Leake, M. Cvetkovic, M. F. Fromm, P. J. Dempsey, M. M. Roden, F. Belas, A. K. Chaudhary, D. M. Roden, A. J. Wood and G. R. Wilkinson (1999). "Interrelationship between substrates and inhibitors of human CYP3A and P-glycoprotein." Pharmaceutical research **16**(3): 408-414.

Kim, Y. and R. Nielsen (2004). "Linkage disequilibrium as a signature of selective sweeps." Genetics **167**(3): 1513-1524.

Kimura, M. (1979). "The neutral theory of molecular evolution." Scientific American **241**(5): 98-100, 102, 108 passim.

Kimura, M. (1991). "The neutral theory of molecular evolution: a review of recent evidence." Idengaku zasshi **66**(4): 367-386.

Kimura, M. and T. Ohta (1978). "Stepwise mutation model and distribution of allelic frequencies in a finite population." Proceedings of the National Academy of Sciences of the United States of America **75**(6): 2868-2872.

Kirchheiner, J. and A. Seeringer (2007). "Clinical implications of pharmacogenetics of cytochrome P450 drug metabolizing enzymes." Biochim Biophys Acta **1770**(3): 489-494.

Kitano, T., Y. H. Liu, S. Ueda and N. Saitou (2004). "Human-specific amino acid changes found in 103 protein-coding genes." Molecular biology and evolution **21**(5): 936-944.

Kivisild, T., M. Reidla, E. Metspalu, A. Rosa, A. Brehm, E. Pennarun, J. Parik, T. Geberhiwot, E. Usanga and R. Villems (2004). "Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears." Am J Hum Genet **75**(5): 752-770.

Kohlrausch, F. B., C. S. Gama, M. I. Lobato, P. Belmonte-de-Abreu, S. M. Callegari-Jacques, A. Gesteira, F. Barros, A. Carracedo and M. H. Hutz (2008). "Naturalistic pharmacogenetic study of treatment resistance to typical neuroleptics in European-Brazilian schizophrenics." Pharmacogenet Genomics **18**(7): 599-609.

Kolchanov, N. A., E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, I. L. Stepanenko, T. I. Merkulova, M. A. Pozdnyakov, N. L. Podkolodny, A. N. Naumochkin and A. G. Romashchenko (2002).

"Transcription Regulatory Regions Database (TRRD): its status in 2002." Nucleic acids research **30**(1): 312-317.

Kolell, K. J. and D. L. Crawford (2002). "Evolution of Sp transcription factors." Molecular biology and evolution **19**(3): 216-222.

Kuehl, P., J. Zhang, Y. Lin, J. Lamba, M. Assem, J. Schuetz, P. B. Watkins, A. Daly, S. A. Wrighton, S. D. Hall, P. Maurel, M. Relling, C. Brimer, K. Yasuda, R. Venkataramanan, S. Strom, K. Thummel, M. S. Boguski and E. Schuetz (2001). "Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression." Nat Genet **27**(4): 383-391.

Kuehn, B. M. (2008). "1000 Genomes Project promises closer look at variation in human genome." JAMA : the journal of the American Medical Association **300**(23): 2715.

Ladunga, I. (2010). Computational biology of transcription factor binding. New York, NY, Humana Press.

Lamba, J. K., Y. S. Lin, E. G. Schuetz and K. E. Thummel (2002). "Genetic contribution to variable human CYP3A-mediated metabolism." Adv Drug Deliv Rev **54**(10): 1271-1294.

Lapidot, M., O. Mizrahi-Man and Y. Pilpel (2008). "Functional characterization of variations on regulatory motifs." PLoS genetics **4**(3): e1000018.

Lee, A. C., A. Kamalam, S. M. Adams and M. A. Jobling (2004). "Molecular evidence for absence of Y-linkage of the Hairy Ears trait." European journal of human genetics : EJHG **12**(12): 1077-1079.

Lee, S. J., K. A. Usmani, B. Chanas, B. Ghanayem, T. Xi, E. Hodgson, H. W. Mohrenweiser and J. A. Goldstein (2003). "Genetic findings and functional studies of human CYP3A5 single nucleotide polymorphisms in different ethnic groups." Pharmacogenetics **13**(8): 461-472.

Lewis, I. M. (1994). Peoples of the Horn of Africa : Somali, Afar and Saho. London, HAAN.

Lewis, M. P. and SIL International (2009). Ethnologue : languages of the world. Dallas, Tex., SIL International.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.

Li, J., L. Zhang, H. Zhou, M. Stoneking and K. Tang (2011). "Global patterns of genetic diversity and signals of natural selection for human ADME genes." Human molecular genetics **20**(3): 528-540.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza and R. M. Myers (2008). "Worldwide human relationships inferred from genome-wide patterns of variation." Science **319**(5866): 1100-1104.

Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**(11): 1451-1452.

Lin, Y. S., A. L. Dowling, S. D. Quigley, F. M. Farin, J. Zhang, J. Lamba, E. G. Schuetz and K. E. Thummel (2002). "Co-regulation of CYP3A4 and CYP3A5 and contribution to hepatic and intestinal midazolam metabolism." Mol Pharmacol **62**(1): 162-172.

Liu, K. and S. V. Muse (2005). "PowerMarker: an integrated analysis environment for genetic marker analysis." Bioinformatics **21**(9): 2128-2129.

Lomelin, D., E. Jorgenson and N. Risch (2010). "Human genetic variation recognizes functional elements in noncoding sequence." Genome research **20**(3): 311-319.

Lovell, A., C. Moreau, V. Yotova, F. Xiao, S. Bourgeois, D. Gehl, J. Bertranpetit, E. Schurr and D. Labuda (2005). "Ethiopia: between Sub-Saharan Africa and western Eurasia." Ann Hum Genet **69**(Pt 3): 275-287.

Lyamichev, V., M. A. Brow and J. E. Dahlberg (1993). "Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases." Science **260**(5109): 778-783.

Lynch, T. and A. Price (2007). "The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects." Am Fam Physician **76**(3): 391-396.

Mabayoje, J. O. (1956). "Sickle-cell anaemia; a major disease in West Africa." Br Med J **1**(4960): 194-196.

MacArthur, R. D. and R. M. Novak (2008). "Reviews of anti-infective agents: maraviroc: the first of a new class of antiretroviral agents." Clin Infect Dis **47**(2): 236-241.

Manica, A., W. Amos, F. Balloux and T. Hanihara (2007). "The effect of ancient population bottlenecks on human phenotypic variation." Nature **448**(7151): 346-348.

Mantel, N. (1967). "The detection of disease clustering and a generalized regression approach." Cancer research **27**(2): 209-220.

Mardis, E. R. (2006). "Anticipating the 1,000 dollar genome." Genome biology **7**(7): 112.

Martin, C., M. Chevrot, H. Poirier, P. Passilly-Degrace, I. Niot and P. Besnard (2011). "CD36 as a lipid sensor." Physiology & behavior **105**(1): 36-42.

Martinez-Jimenez, C. P., M. J. Gomez-Lechon, J. V. Castell and R. Jover (2005). "Transcriptional regulation of the human hepatic CYP3A4: identification of a new distal enhancer region responsive to CCAAT/enhancer-binding protein beta isoforms (liver activating protein and liver inhibitory protein)." Molecular pharmacology **67**(6): 2088-2101.

Martins, E. P. and E. A. Housworth (2002). "Phylogeny shape and the phylogenetic comparative method." Systematic biology **51**(6): 873-880.

Matsumura, K., T. Saito, Y. Takahashi, T. Ozeki, K. Kiyotani, M. Fujieda, H. Yamazaki, H. Kunitoh and T. Kamataki (2004). "Identification of a novel polymorphic enhancer of the human CYP3A4 gene." Molecular pharmacology **65**(2): 326-334.

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

McDougall, I., F. H. Brown and J. G. Fleagle (2005). "Stratigraphic placement and age of modern humans from Kibish, Ethiopia." Nature **433**(7027): 733-736.

McGuigan, F. E. and S. H. Ralston (2002). "Single nucleotide polymorphism detection: allelic discrimination using TaqMan." Psychiatr Genet **12**(3): 133-136.

McGuire, A. L. (2008). "1000 Genomes on the Road to Personalized Medicine." Personalized medicine **5**(3): 195-197.

Mirghani, R. A., J. Sayi, E. Aklillu, A. Allqvist, M. Jande, A. Wennerholm, J. Eriksen, V. M. Herben, B. C. Jones, L. L. Gustafsson and L. Bertilsson (2006). "CYP3A5 genotype has significant effect on quinine 3-hydroxylation in Tanzanians, who have lower total CYP3A activity than a Swedish population." Pharmacogenet Genomics **16**(9): 637-645.

Moilanen, A. M., J. Hakkola, M. H. Vaarala, S. Kauppila, P. Hirvikoski, J. T. Vuoristo, R. J. Edwards and T. K. Paavonen (2007). "Characterization of androgen-regulated expression of CYP3A5 in human prostate." Carcinogenesis **28**(5): 916-921.

Mouly, S. J., C. Matheny, M. F. Paine, G. Smith, J. Lamba, V. Lamba, S. N. Pusek, E. G. Schuetz, P. W. Stewart and P. B. Watkins (2005). "Variation in oral clearance of saquinavir is predicted by CYP3A5*1 genotype but not by enterocyte content of cytochrome P450 3A5." Clin Pharmacol Ther **78**(6): 605-618.

Mukonzo, J. K., P. Waako, J. Ogwal-Okeng, L. L. Gustafsson and E. Aklillu "Genetic variations in ABCB1 and CYP3A5 as well as sex influence quinine disposition among Ugandans." Ther Drug Monit **32**(3): 346-352.

Murdock, G. P. (1959). Africa: its peoples and their culture history. [With maps and illustrations.], pp. xiii. 456. McGraw-Hill Book Co.: New York**:** 8º.

Myerowitz, R. (2001). "The search for the genetic lesion in Ashkenazi Jews with Classic Tay-Sachs disease." Adv Genet **44**: 137-143.

Nachman, M. W. and S. L. Crowell (2000). "Estimate of the mutation rate per nucleotide in humans." Genetics **156**(1): 297-304.

Nakamura, T., Y. Saito, N. Murayama, M. Saeki, A. Soyama, K. Sai, S. Ozawa and J. Sawada (2001). "Apparent low frequency of sequence variability within the proximal promoter region of the cytochrome P450 (CYP) 3A5 gene in established cell lines from Japanese individuals." Biological & pharmaceutical bulletin **24**(8): 954-957.

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

Nei, M. (1987). Molecular evolutionary genetics. New York, Columbia University Press.

Nei, M. and S. Kumar (2000). Molecular evolution and phylogenetics. New York, Oxford University Press.

Neilson, J. R. and R. Sandberg (2010). "Heterogeneity in mammalian RNA 3' end formation." Experimental cell research **316**(8): 1357-1364.

Nelson, D. R. (2009). "The cytochrome p450 homepage." Hum Genomics **4**(1): 59-65.

Nelson, D. R., D. C. Zeldin, S. M. Hoffman, L. J. Maltais, H. M. Wain and D. W. Nebert (2004). "Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants." Pharmacogenetics **14**(1): 1-18.

Nielsen, R., M. J. Hubisz, I. Hellmann, D. Torgerson, A. M. Andres, A. Albrechtsen, R. Gutenkunst, M. D. Adams, M. Cargill, A. Boyko, A. Indap, C. D. Bustamante and A. G. Clark (2009). "Darwinian and demographic forces affecting human protein coding genes." Genome research **19**(5): 838-849.

Oliveira, E., R. Pereira, A. Amorim, H. McLeod and M. J. Prata (2009). "Patterns of pharmacogenetic diversity in African populations: role of ancient and recent history." Pharmacogenomics **10**(9): 1413-1422.

Pain, A., B. C. Urban, O. Kai, C. Casals-Pascual, J. Shafi, K. Marsh and D. J. Roberts (2001). "A non-sense mutation in Cd36 gene is associated with protection from severe malaria." Lancet **357**(9267): 1502-1503.

Pakenham, T. (1991). The scramble for Africa, 1876-1912. London, Weidenfeld and Nicolson.

Parra, E. J., A. Marcini, J. Akey, J. Martinson, M. A. Batzer, R. Cooper, T. Forrester, D. B. Allison, R. Deka, R. E. Ferrell and M. D. Shriver (1998). "Estimating African American admixture proportions by use of population-specific alleles." Am J Hum Genet **63**(6): 1839-1851.

Patki, K. C., L. L. Von Moltke and D. J. Greenblatt (2003). "In vitro metabolism of midazolam, triazolam, nifedipine, and testosterone by human liver microsomes and recombinant cytochromes p450: role of cyp3a4 and cyp3a5." Drug metabolism and disposition: the biological fate of chemicals **31**(7): 938-944.

Patrinos, G. P. and M. B. Petersen (2009). "Copy number variation and genomic alterations in health and disease." Genome Med **1**(2): 21.

Patterson, K. (2011). "1000 genomes: a world of variation." Circulation research **108**(5): 534-536.

Paulussen, A., K. Lavrijsen, H. Bohets, J. Hendrickx, P. Verhasselt, W. Luyten, F. Konings and M. Armstrong (2000). "Two linked mutations in transcriptional regulatory elements of the CYP3A5 gene constitute the major genetic determinant of polymorphic activity in humans." Pharmacogenetics **10**(5): 415-424.

Payseur, B. A., A. D. Cutter and M. W. Nachman (2002). "Searching for evidence of positive selection in the human genome using patterns of microsatellite variability." Molecular biology and evolution **19**(7): 1143-1153.

Pennisi, E. (2010). "Genomics. 1000 Genomes Project gives new map of genetic diversity." Science **330**(6004): 574-575.

Perry, G. H., N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee and A. C. Stone (2007). "Diet and the evolution of human amylase gene copy number variation." Nat Genet **39**(10): 1256-1260.

Piedade, R. and J. P. Gil (2011). "The pharmacogenetics of antimalaria artemisinin combination therapy." Expert opinion on drug metabolism & toxicology **7**(10): 1185-1200.

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

Plummer, S. J., D. V. Conti, P. L. Paris, A. P. Curran, G. Casey and J. S. Witte (2003). "CYP3A4 and CYP3A5 genotypes, haplotypes, and risk of prostate cancer." Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology **12**(9): 928-932.

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

Pritchard, J. K., J. K. Pickrell and G. Coop (2010). "The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation." Current biology : CB **20**(4): R208-215.

Prugnolle, F., A. Manica and F. Balloux (2005). "Geography predicts neutral genetic diversity of human populations." Curr Biol **15**(5): R159-160.

Qiu, H., S. Taudien, H. Herlyn, J. Schmitz, Y. Zhou, G. Chen, R. Roberto, M. Rocchi, M. Platzer and L. Wojnowski (2008). "CYP3 phylogenomics: evidence for positive selection of CYP3A4 and CYP3A7." Pharmacogenetics and genomics **18**(1): 53-66.

Quaranta, S., D. Chevalier, D. Allorge, J. M. Lo-Guidice, F. Migot-Nabias, A. Kenani, M. Imbenotte, F. Broly, B. Lacarelle and M. Lhermitte (2006). "Ethnic differences in the distribution of CYP3A5 gene polymorphisms." Xenobiotica **36**(12): 1191-1200.

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

Quteineh, L., C. Verstuyft, V. Furlan, A. Durrbach, A. Letierce, S. Ferlicot, A. M. Taburet, B. Charpentier and L. Becquemont (2008). "Influence of CYP3A5 genetic polymorphism on tacrolimus daily dose requirements and acute rejection in renal graft recipients." Basic Clin Pharmacol Toxicol **103**(6): 546-552.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman and L. L. Cavalli-Sforza (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**(44): 15942-15947.

Rannala, B. and G. Bertorelle (2001). "Using linked markers to infer the age of a mutation." Human mutation **18**(2): 87-100.

Raymond, M. and F. Rousset (1995). "An exact test for population differentiation." Evolution **49**(6): 1280-1283.

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

Reed, T. E. (1969). "Caucasian genes in American Negroes." Science **165**(895): 762-768.

Reese, M. G., F. H. Eeckman, D. Kulp and D. Haussler (1997). "Improved splice site detection in Genie." Journal of computational biology : a journal of computational molecular cell biology **4**(3): 311-323.

Relethford, J. H. (2008). "Genetic evidence and the modern human origins debate." Heredity **100**(6): 555-563.

Richards, M., C. Rengo, F. Cruciani, F. Gratrix, J. F. Wilson, R. Scozzari, V. Macaulay and A. Torroni (2003). "Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations." Am J Hum Genet **72**(4): 1058-1064.

Ritchie, M. D. and W. S. Bush (2010). "Genome simulation approaches for synthesizing in silico datasets for human genomics." Advances in genetics **72**: 1-24.

Roder, K., S. S. Wolf, K. J. Larkin and M. Schweizer (1999). "Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1." Gene **234**(1): 61-69.

Rodriguez-Antona, C., M. Jande, A. Rane and M. Ingelman-Sundberg (2005). "Identification and phenotype characterization of two CYP3A haplotypes causing different enzymatic capacity in fetal livers." Clinical pharmacology and therapeutics **77**(4): 259-270.

Rotimi, C. N. and L. B. Jorde (2010). "Ancestry and disease in the age of genomic medicine." The New England journal of medicine **363**(16): 1551-1558.

Roy, J. N., J. Lajoie, L. S. Zijenah, A. Barama, C. Poirier, B. J. Ward and M. Roger (2005). "CYP3A5 genetic polymorphisms in different ethnic populations." Drug Metab Dispos **33**(7): 884-887.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward and E. S. Lander (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-837.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler and E. S. Lander (2006). "Positive natural selection in the human lineage." Science **312**(5780): 1614-1620.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L.

Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, P. C. Sham, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, I. Yakub, B. W. Birren, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson and J. Stewart (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-918.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander and D. Altshuler (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-933.

Saeki, M., Y. Saito, T. Nakamura, N. Murayama, S. R. Kim, S. Ozawa, K. Komamura, K. Ueno, S. Kamakura, T. Nakajima, H. Saito, Y. Kitamura, N. Kamatani and J. Sawada (2003). "Single nucleotide polymorphisms and haplotype frequencies of CYP3A5 in a Japanese population." Hum Mutat **21**(6): 653.

Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay and A. Carracedo (2002). "The making of the African mtDNA landscape." American journal of human genetics **71**(5): 1082-1111.

Salkind, N. J. (2007). Encyclopedia of measurement and statistics. Thousand Oaks, Calif. ; London, SAGE Publications.

Salzburger, W., G. B. Ewing and A. Von Haeseler (2011). "The performance of phylogenetic algorithms in estimating haplotype genealogies with migration." Molecular ecology **20**(9): 1952-1963.

Samani, N. J., M. Tomaszewski and H. Schunkert (2010). "The personal genome--the future of personalised medicine?" Lancet **375**(9725): 1497-1498.

Scherer, S. W., C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler, N. P. Carter, M. E. Hurles and L. Feuk (2007). "Challenges and standards in integrating surveys of structural variation." Nat Genet **39**(7 Suppl): S7-15.

Schirmer, M., M. R. Toliat, M. Haberl, A. Suk, L. K. Kamdem, K. Klein, J. Brockmoller, P. Nurnberg, U. M. Zanger and L. Wojnowski (2006). "Genetic signature consistent with selection against the CYP3A4*1B allele in non-African populations." Pharmacogenet Genomics **16**(1): 59-71.

Schmitz, A., S. Demmel, L. M. Peters, T. Leeb, M. Mevissen and B. Haase (2010). "Comparative human-horse sequence analysis of the CYP3A subfamily gene cluster." Animal genetics **41 Suppl 2**: 72-79.

Scholz, C. A., T. C. Johnson, A. S. Cohen, J. W. King, J. A. Peck, J. T. Overpeck, M. R. Talbot, E. T. Brown, L. Kalindekafe, P. Y. Amoako, R. P. Lyons, T. M. Shanahan, I. S. Castaneda, C. W. Heil, S. L. Forman, L. R. McHargue, K. R. Beuning, J. Gomez and J. Pierson (2007). "East African megadroughts between 135 and 75 thousand years ago and bearing on early-modern human origins." Proc Natl Acad Sci U S A **104**(42): 16416-16421.

Schuetz, J. D., D. T. Molowa and P. S. Guzelian (1989). "Characterization of a cDNA encoding a new member of the glucocorticoid-responsive cytochromes P450 in human liver." Arch Biochem Biophys **274**(2): 355-365.

Schuster, S. C., W. Miller, A. Ratan, L. P. Tomsho, B. Giardine, L. R. Kasson, R. S. Harris, D. C. Petersen, F. Zhao, J. Qi, C. Alkan, J. M. Kidd, Y. Sun, D. I. Drautz, P. Bouffard, D. M. Muzny, J. G. Reid, L. V. Nazareth, Q. Wang, R. Burhans, C. Riemer, N. E. Wittekindt, P. Moorjani, E. A. Tindall, C. G. Danko, W. S. Teo, A. M. Buboltz, Z. Zhang, Q. Ma, A. Oosthuysen, A. W. Steenkamp, H. Oostuisen, P. Venter, J. Gajewski, Y. Zhang, B. F. Pugh, K. D. Makova, A. Nekrutenko, E. R. Mardis, N. Patterson, T. H. Pringle, F. Chiaromonte, J. C. Mullikin, E. E. Eichler, R. C. Hardison, R. A. Gibbs, T. T. Harkins and V. M. Hayes (2010). "Complete Khoisan and Bantu genomes from southern Africa." Nature **463**(7283): 943-947.

Service, R. F. (2006). "Gene sequencing. The race for the $1000 genome." Science **311**(5767): 1544-1546.

Sgaier, S. K., P. Jha, P. Mony, A. Kurpad, V. Lakshmi, R. Kumar and N. K. Ganguly (2007). "Public health. Biobanks in developing countries: needs and feasibility." Science **318**(5853): 1074-1075.

Shapiro, M. B. and P. Senapathy (1987). "RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression." Nucleic Acids Res **15**(17): 7155-7174.

Shen, G. Q., K. G. Abdullah and Q. K. Wang (2009). "The TaqMan method for SNP genotyping." Methods Mol Biol **578**: 293-306.

Shreeve, J. (1994). "'Lucy,' crucial early human ancestor, finally gets a head." Science **264**(5155): 34-35.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro (1995). "Properties of statistical tests of neutrality for DNA polymorphism data." Genetics **141**(1): 413-429.

Sinues, B., J. Vicente, A. Fanlo, E. Mayayo-Sinues, F. Gonzalez-Andrade, Q. D. Sanchez and B. Martinez-Jarreta (2008). "CYP3A5 3, CYP3A4 1B and MDR1 C3435T genotype distributions in Ecuadorians." Dis Markers **24**(6): 325-331.

Sinues, B., J. Vicente, A. Fanlo, P. Vasquez, J. C. Medina, E. Mayayo, B. Conde, I. Arenaz and B. Martinez-Jarreta (2007). "CYP3A5*3 and CYP3A4*1B allele distribution and genotype combinations: differences between Spaniards and Central Americans." Ther Drug Monit **29**(4): 412-416.

Siva, N. (2008). "1000 Genomes project." Nature biotechnology **26**(3): 256.

Slatkin, M. (1995). "A measure of population subdivision based on microsatellite allele frequencies." Genetics **139**(1): 457-462.

Slatkin, M. and B. Rannala (2000). "Estimating allele age." Annual review of genomics and human genetics **1**: 225-249.

Slatko, B. E., L. M. Albright, S. Tabor and J. Ju (2001). "DNA sequencing by the dideoxy method." Curr Protoc Mol Biol **Chapter 7**: Unit7 4A.

Smits, H. L. (2009). "Prospects for the control of neglected tropical diseases by mass drug administration." Expert Rev Anti Infect Ther **7**(1): 37-56.

Sobngwi, E., J. C. Mbanya, N. C. Unwin, A. P. Kengne, L. Fezeu, E. M. Minkoulou, T. J. Aspray and K. G. Alberti (2002). "Physical activity and its relationship with obesity, hypertension and diabetes in urban and rural Cameroon." International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity **26**(7): 1009-1016.

Solomon, A. W., S. Nayagam and G. Pasvol (2009). "Recent advances in tropical medicine." Trans R Soc Trop Med Hyg **103**(7): 647-652.

Stephens, M. and P. Donnelly (2003). "A comparison of bayesian methods for haplotype reconstruction from population genotype data." Am J Hum Genet **73**(5): 1162-1169.

Stephens, M., N. J. Smith and P. Donnelly (2001). "A new statistical method for haplotype reconstruction from population data." Am J Hum Genet **68**(4): 978-989.

Suske, G. (1999). "The Sp-family of transcription factors." Gene **238**(2): 291-300.

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

Tamura, K., J. Dudley, M. Nei and S. Kumar (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Molecular biology and evolution **24**(8): 1596-1599.

Thomas, M. G., N. Bradman and H. M. Flinn (1999). "High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome." Human genetics **105**(6): 577-581.

Thomas, M. G., M. E. Weale, A. L. Jones, M. Richards, A. Smith, N. Redhead, A. Torroni, R. Scozzari, F. Gratrix, A. Tarekegn, J. F. Wilson, C. Capelli, N. Bradman and D. B. Goldstein (2002). "Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors." American journal of human genetics **70**(6): 1411-1420.

Thompson, E. E., H. Kuttab-Boulos, D. Witonsky, L. Yang, B. A. Roe and A. Di Rienzo (2004). "CYP3A variation and the evolution of salt-sensitivity variants." Am J Hum Genet **75**(6): 1059-1069.

Thompson, E. E., H. Kuttab-Boulos, L. Yang, B. A. Roe and A. Di Rienzo (2006). "Sequence diversity and haplotype structure at the human CYP3A cluster." Pharmacogenomics J **6**(2): 105-114.

Tian, B., J. Hu, H. Zhang and C. S. Lutz (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." Nucleic acids research **33**(1): 201-212.

Tintle, N., H. Aschard, I. Hu, N. Nock, H. Wang and E. Pugh (2011). "Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17." Genetic epidemiology **35 Suppl 1**: S56-60.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber and S. M. Williams (2009). "The genetic structure and history of Africans and African Americans." Science **324**(5930): 1035-1044.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray and P. Deloukas (2007). "Convergent adaptation of human lactase persistence in Africa and Europe." Nat Genet **39**(1): 31-40.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E. H. Touma, S. M. Williams and A. G. Clark (2001). "Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance." Science **293**(5529): 455-462.

Tishkoff, S. A. and B. C. Verrelli (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 S. M. Williams (2002). "Genetic analysis of African populations: human evolution and complex disease." Nat Rev Genet **3**(8): 611-621.

Tremblay, M. and H. Vezina (2000). "New estimates of intergenerational time intervals for the calculation of age and origins of mutations." American journal of human genetics **66**(2): 651-658.

Trowell, H. C., A. B. Raper and H. F. Welbourn (1957). "The natural history of homozygous sickle-cell anaemia in Central Africa." Q J Med **26**(104): 401-422.

Veeramah, K. R., M. G. Thomas, M. E. Weale, D. Zeitlyn, A. Tarekegn, E. Bekele, N. R. Mendell, E. A. Shephard, N. Bradman and I. R. Phillips (2008). "The potentially deleterious functional variant

flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa." Pharmacogenet Genomics **18**(10): 877-886.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard (2006). "A map of recent positive selection in the human genome." PLoS biology **4**(3): e72.

Wade, N. (2006). "The quest for the $1,000 human genome: DNA sequencing in the doctor's office? At birth? It may be coming closer." The New York times: F1, F3.

Wain, L. V., J. A. Armour and M. D. Tobin (2009). "Genomic copy number variation, human health, and disease." Lancet **374**(9686): 340-350.

Walker, R., N. Unwin and K. G. Alberti (1998). "Hypertension treatment and control in Sub-saharan Africa. Burden of cerebrovascular disease will increase as more people survive to old age." BMJ **317**(7150): 76-77; author reply 77.

Wang, E. T., G. Kodama, P. Baldi and R. K. Moyzis (2006). "Global landscape of recent inferred Darwinian selection for Homo sapiens." Proceedings of the National Academy of Sciences of the United States of America **103**(1): 135-140.

Watkins, W. S., C. E. Ricker, M. J. Bamshad, M. L. Carroll, S. V. Nguyen, M. A. Batzer, H. C. Harpending, A. R. Rogers and L. B. Jorde (2001). "Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms." Am J Hum Genet **68**(3): 738-752.

Watlington, C. O., L. B. Kramer, E. G. Schuetz, J. Zilai, W. M. Grogan, P. Guzelian, F. Gizek and A. C. Schoolwerth (1992). "Corticosterone 6 beta-hydroxylation correlates with blood pressure in spontaneously hypertensive rats." Am J Physiol **262**(6 Pt 2): F927-931.

Webb, K. E., J. F. Martin, J. Cotton, J. D. Erusalimsky and S. E. Humphries (2003). "The 4830C>A polymorphism within intron 5 affects the pattern of alternative splicing occurring within exon 6 of the thrombopoietin gene." Experimental hematology **31**(6): 488-494.

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

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

Whittaker, J. C., R. M. Harbord, N. Boxall, I. Mackay, G. Dawson and R. M. Sibly (2003). "Likelihood-based estimation of microsatellite mutation rates." Genetics **164**(2): 781-787.

Wilke, R. A., D. W. Lin, D. M. Roden, P. B. Watkins, D. Flockhart, I. Zineh, K. M. Giacomini and R. M. Krauss (2007). "Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges." Nat Rev Drug Discov **6**(11): 904-916.

Williams, E. T., A. S. Rodin and H. W. Strobel (2004). "Defining relationships between the known members of the cytochrome P450 3A subfamily, including five putative chimpanzee members." Mol Phylogenet Evol **33**(2): 300-308.

Williams, E. T., K. R. Schouest, M. Leyk and H. W. Strobel (2007). "The chimpanzee cytochrome P450 3A subfamily: Is our closest related species really that similar?" Comp Biochem Physiol Part D Genomics Proteomics **2**(2): 91-100.

Williams, P. A., J. Cosme, D. M. Vinkovic, A. Ward, H. C. Angove, P. J. Day, C. Vonrhein, I. J. Tickle and H. Jhoti (2004). "Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone." Science **305**(5684): 683-686.

Wilson, I. J. and D. J. Balding (1998). "Genealogical inference from microsatellite data." Genetics **150**(1): 499-510.

Wojnowski, L. (2004). "Genetics of the variable expression of CYP3A in humans." Ther Drug Monit **26**(2): 192-199.

Wojnowski, L., P. C. Turner, B. Pedersen, E. Hustert, J. Brockmoller, M. Mendy, H. C. Whittle, G. Kirk and C. P. Wild (2004). "Increased levels of aflatoxin-albumin adducts are associated with CYP3A5 polymorphisms in The Gambia, West Africa." Pharmacogenetics **14**(10): 691-700.

Wolf-Maier, K., R. S. Cooper, J. R. Banegas, S. Giampaoli, H. W. Hense, M. Joffres, M. Kastarinen, N. Poulter, P. Primatesta, F. Rodriguez-Artalejo, B. Stegmayr, M. Thamm, J. Tuomilehto, D. Vanuzzo and F. Vescio (2003). "Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States." JAMA **289**(18): 2363-2369.

Wong, M., R. L. Balleine, M. Collins, C. Liddle, C. L. Clarke and H. Gurney (2004). "CYP3A5 genotype and midazolam clearance in Australian patients receiving chemotherapy." Clin Pharmacol Ther **75**(6): 529-538.

Wood, E. T., D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, L. Louie, M. Bamshad, B. I. Strassmann, H. Soodyall and M. F. Hammer (2005). "Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes." European journal of human genetics : EJHG **13**(7): 867-876.

Wright, S. (1950). "Genetic structure of populations." British medical journal **2**(4669): 36.

Wright, S. I. and B. Charlesworth (2004). "The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model." Genetics **168**(2): 1071-1076.

Wrighton, S. A., W. R. Brian, M. A. Sari, M. Iwasaki, F. P. Guengerich, J. L. Raucy, D. T. Molowa and M. Vandenbranden (1990). "Studies on the expression and metabolic capabilities of human liver cytochrome P450IIIA5 (HLp3)." Mol Pharmacol **38**(2): 207-213.

Wurthwein, R., A. Gbangou, R. Sauerborn and C. M. Schmidt (2001). "Measuring the local burden of disease. A study of years of life lost in sub-Saharan Africa." International journal of epidemiology **30**(3): 501-508.

Xie, H. G., A. J. Wood, R. B. Kim, C. M. Stein and G. R. Wilkinson (2004). "Genetic variability in CYP3A5 and its possible consequences." Pharmacogenomics **5**(3): 243-272.

Xu, C. F., K. Lewis, K. L. Cantone, P. Khan, C. Donnelly, N. White, N. Crocker, P. R. Boyd, D. V. Zaykin and I. J. Purvis (2002). "Effectiveness of computational methods in haplotype prediction." Human genetics **110**(2): 148-156.

Xu, H., R. Chakraborty and Y. X. Fu (2005). "Mutation rate variation at human dinucleotide microsatellites." Genetics **170**(1): 305-312.

Yamaori, S., H. Yamazaki, S. Iwano, K. Kiyotani, K. Matsumura, T. Saito, A. Parkinson, K. Nakagawa and T. Kamataki (2005). "Ethnic differences between Japanese and Caucasians in the expression levels of mRNAs for CYP3A4, CYP3A5 and CYP3A7: lack of co-regulation of the expression of CYP3A in Japanese livers." Xenobiotica **35**(1): 69-83.

Young, J. H., Y. P. Chang, J. D. Kim, J. P. Chretien, M. J. Klag, M. A. Levine, C. B. Ruff, N. Y. Wang and A. Chakravarti (2005). "Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion." PLoS genetics **1**(6): e82.

Yu, K. S., J. Y. Cho, I. J. Jang, K. S. Hong, J. Y. Chung, J. R. Kim, H. S. Lim, D. S. Oh, S. Y. Yi, K. H. Liu, J. G. Shin and S. G. Shin (2004). "Effect of the CYP3A5 genotype on the pharmacokinetics of intravenous midazolam during inhibited and induced metabolic states." Clin Pharmacol Ther **76**(2): 104-112.

Zaphiropoulos, P. G. (2003). "A map of the mouse Cyp3a locus." DNA Seq **14**(3): 155-162.

Zhang, W. and M. E. Dolan (2010). "Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery." Pharmacogenomics **11**(2): 249-256.

Zhao, Y., M. Song, D. Guan, S. Bi, J. Meng, Q. Li and W. Wang (2005). "Genetic polymorphisms of CYP3A5 genes and concentration of the cyclosporine and tacrolimus." Transplant Proc **37**(1): 178-181.

Zheng, H., S. Webber, A. Zeevi, E. Schuetz, J. Zhang, P. Bowman, G. Boyle, Y. Law, S. Miller, J. Lamba and G. J. Burckart (2003). "Tacrolimus dosing in pediatric heart transplant patients is related to CYP3A5 and MDR1 gene polymorphisms." Am J Transplant **3**(4): 477-483.

Zhenhua, L., N. Tsuchiya, S. Narita, T. Inoue, Y. Horikawa, H. Kakinuma, T. Kato, O. Ogawa and T. Habuchi (2005). "CYP3A5 gene polymorphism and risk of prostate cancer in a Japanese population." Cancer Lett **225**(2): 237-243.