Genetic variation in the FMO2 gene: evolution & functional consequences

Al-Sulaimani, Maha Saleh

Genetic Variation in the *FMO2* Gene: Evolution & Functional Consequences

Maha Saleh Al-Sulaimani

School of Biological and Chemical Sciences

Queen Mary, University of London

Submitted for the degree of Doctor of Philosophy

Supervisor: Prof. Ian R. Phillips

Declaration of Ownership

I, Maha Saleh Al-Sulaimani, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Flavin-containing monooxygenase 2 (FMO2) is involved in the metabolism of xenobiotics, including therapeutic drugs. FMO2 exists in two forms: a functional and a non-functional form. The functional allele is found only in Africa and individuals of recent African origin. The aims of the project were to determine the frequency of functional FMO2 in Africa and obtain insights into the evolutionary history of the *FMO2* gene.

Six hundred and eighty nine samples from nine African population groups were genotyped for six high-frequency SNPs, and the genetic diversity within *FMO2* was characterized by sequencing 3.44 kb of genomic DNA, encompassing the entire coding sequence and some flanking intronic sequences in 48 African individuals. Haplotypes were inferred using Phase and the relationship between mutations was revealed using reduced-median and median-joining Network. Test statistics were used to determine whether the genetic variation is compatible with neutral evolution.

Genotyping indicated that deleterious SNPs occur mostly on a non-functional allele and that the frequencies of three were significantly different (P<0.05) among populations. Resequencing identified 32 variants. Genetree was used to estimate the time to the most recent common ancestral sequence (~0.928 million years) and the ages of some of the mutations.

Results indicate that the frequency of full-length 23238C alleles is relatively uniform across sub-Saharan Africa. Interestingly, this is not the case for the inferred potentially functional 23238C alleles, which frequency differed significantly (P<0.05) across sub-Saharan Africa.

The results also provide evidence that the frequency of functional FMO2 in east and west-Africa is high (≥0.54), which has important implications for therapy with drugs that are substrates for FMO2.

A $K_a/K_s > 1$, and low nucleotide sequence diversity of intronic regions of 23238C alleles indicate a possible selective sweep.

Abbreviations

ABI, Applied Biosystems

CEPH, Centre d`Etude du Polymorphisme Humain

DNA, Deoxy Ribonucleic Acid

DnaSP, DNA Sequence Polymorphism

ETA, Ethionamide

EGP, Environmental Gene Project

ELB, Estimation Likelihood Bayesian

EM, Estimation Maximization

FMO2, Flavin-containing Monooxygenase 2

FAD-OOH, 4a-Hydroxy Flavin

FEL, Fixed Effect-Likelihood

HapMap-HCB, HapMap Han Chinese Panel

HapMap-Ceu, HapMap European Panel from Utah

HapMap-JPT, HapMap Japanese (Tokyo) Panel

HGDP, Human Genome Diversity Panel

HWE, Hardy-Weinberg Equilibrium

INDEL, Insertion-Deletion Polymorphism

KYR, Hundred Thousand Years

LD, Linkage Disequilibrium

MJ, Median-Joining

MKT, McDonald-Kreitman Test

MYR, Million Years

NCBI, National Centre for Biotechnology Information

NIEHS, National Institute of Environmental Health Sciences

PCR, Polymerase Chain Reaction

REL, Random Effect-Likelihood

RM, Reduced-Median

RNA, Ribunucleic Acid

SMOGD, Software for the Measurement of Genetic Diversity

SNP, Single-Nucleotide Polymorphism

SSCP, Single Strand Conformation Polymorphism

STRPs, Short Tandem Repeat Polymorphisms

TAZ, Thioacetazone

TCGA, The Centre of Genetic Anthropology

TMA, TMAU, Trimethylamine and Trimethylaminuria

T_{MRCA}, Time to the Most Recent Common Ancestor

UCL, University College London

UK, United Kingdom

USA, United States of America

Acknowledgements

I owe special thanks to King Saud University of Riyadh, Saudi Arabia for funding me, and to all individuals involved with donating and collecting samples, especially Dr Ayele Tarekegn, Dr Krishna Veeramah and Dr Sarah Browning.

I would like to thank my supervisor Professor Ian Phillips for his help and support during the duration of this PhD. He has provided a lot of valuable advice which I highly appreciate. I am also grateful for all the guidance that he gave me along the way. Thanks also go to my advisory panel, Professor Richard Nichols and Dr Steve Le Comber for their valuable advice and continuous guidance. I am grateful to Dr Rosemary Ekong for all her help and support in getting me started with genotyping and for all her advice and kind words. I would also like to thank Dr Neil Bradman for providing the opportunity to use the samples available at the TCGA for my research, as well as giving me some useful guidelines. I also appreciate the help of Dr Sarah Browning, Olivia Creemer, Ripu Bairns, Chris Plaster, Dr Krishna Veeramah and Dr Kate Ingram.

Thank you to all at Queen Mary and UCL, who have helped me.

I will be eternally grateful to my amazingly patient and caring husband Majed, for his selfless support and encouragement throughout this journey, all of which have been a motivating power for me, without him this would not have been possible. He was always there ready to advise and cheer me up when I needed it. To our children, Khalid, Deema and Dania, thank you for making me laugh when I least felt like it.

Last but not least, thanks go out to my parents, Saleh (Assoc Prof, Consultant Surgeon, Dr Med Ret) and Ingeborg Al-Sulaimani for teaching me that education is of the utmost importance in any person`s life.

Page of Contents

List of Tables

List of Figures

1. Introduction

1.1 FMOs

Flavin-containing monooxygenases (FMOs; EC 1.14.13.8) were first described by Dr Daniel Ziegler and colleagues in 1964 (reviewed in [Ziegler 1988; Ziegler 1993; Poulsen and Ziegler 1995; Ziegler 2002]). In 1972 an FMO enzyme was isolated from pig liver and characterized by Carolyn Mitchell. The enzyme was first termed "dimethylaniline *N*-oxidase", or "mixed-function amine oxidase" but, with the realization that it could catalyze the oxygenation of a range of substrates, it becomes known as "multisubstrate FMO". FMOs have recently been recognized as belonging to the larger family of Baeyer-Villiger monooxygenases (Krueger and Williams 2005).

FMOs catalyze the oxygenation of drugs and other non-nutritive compounds, including tranquilizers, analgesics and antidepressants, that contain a soft nucleophile, usually nitrogen, sulfur, selenium or phosphorus (Phillips *et al*. 2007; Hao *et al*. 2009). Like cytochromes P450 (CYPs), FMOs utilize the reducing power of NADPH to reduce one atom of molecular oxygen to water, while the other is used to oxidize the substrate (Ziegler and Petit 1964; Krueger and Williams 2005).

FMOs and CYPs also exhibit similar cellular and tissue location, substrate specificity and molecular mass, and exist as multiple enzymes under developmental control. The human FMO gene family is much smaller (one family with five functional members) than that of CYP. The mammalian FMO gene family contains five genes (*FMO1* through *FMO5*) (Lawton *et al*. 1994; Phillips *et al.* 1995). An interesting feature of the enzyme is that substrate binding has no effect on the velocity of the enzyme-catalyzed reaction, since the rate limiting steps of this reaction occur independent of substrate (Ziegler 1988; Krueger and Williams 2005).

It is thought that, over the course of evolutionary history, an arms race has developed between plants and the animals that consume them. To deter animals from consuming them, plants often produce toxic substances, usually through plant secondary metabolism. In turn, animals have evolved both avoidance mechanisms such as bitter taste receptors (many plant toxins are bitter), and metabolic solutions such as detoxifying enzymes like FMO, to circumvent toxicity. As a result, FMO, as with CYP monooxygenases, developed broad substrate specificity at the expense of turnover rate (Krueger and Williams 2005).

In general, the metabolites produced by FMO oxygenation usually have reduced pharmacological and toxicological properties. Sometimes, however, FMO converts chemicals into products that can cause toxicity like in the case of thiourea and if they can be bio-activated by CYP to a toxic metabolite, e.g, an epoxide (Cashman and Zhang 2006; Zhang *et al.* 2006a).

Genetic variability and splicing variants may contribute to the interindividual and interethnic variability observed for metabolism mediated by FMO (Cashman and Zhang 2006).

FMO expression can by regulated by various physiological factors including cofactor supply, diet, insulin (Borbas 2006) and sex hormones (Hukkanen and Dempsey 2005), and this may have implications for the clinical significance of FMO (Cashman *et al.* 1997). It is possible that ancestral FMO was important in processing environmental products (e.g, pesticides), because many such materials are detoxified by FMO (Phillips *et al.* 2007). For example, certain FMOs may have evolved to detoxify specific toxins. The prevalence of abnormal FMO3 in individuals from the tropics and the role of this enzyme in trimethylamine (TMA) metabolism may be an example of evolutionary pressure to decrease metabolism of TMA so that TMA could be used as some primitive volatile insecticide (Cashman *et al.* 1997; Cashman and Zhang 2006).

1.2 Human FMO gene family

To date, eleven FMO genes have been discovered, five of which are functional, designated *FMO1-FMO5*, whereas the others, *FMO6P-FMO11P*, are pseudogenes (Hernandez *et al*. 2004). *FMO1* to *FMO4* lie within a 220-kb cluster on chromosome 1, in the region q24.3. *FMO5* is located ~26 Mb closer to the centromere at 1q21.1. A sixth gene, which is present in the cluster, is *FMO6P*. Five other FMO genes which are all pseudogenes, *FMO7P* to *FMO11P*, are located within a cluster ~4 Mb closer to the centromere than the functional FMO gene cluster (Hernandez *et al*. 2004) (Figure 1). Interestingly, 3 of these pseodogenes are functional in mice (Hernandez *et al*. 2004).

It is predicted that all mammals possess the functional FMO genes *FMO1* to *FMO5*, since phylogenetic analysis suggests that these genes arose from a common ancestor gene through a series of gene duplications which predated the divergence of mammals, some 85 million years ago (Hernandez *et al*. 2004).

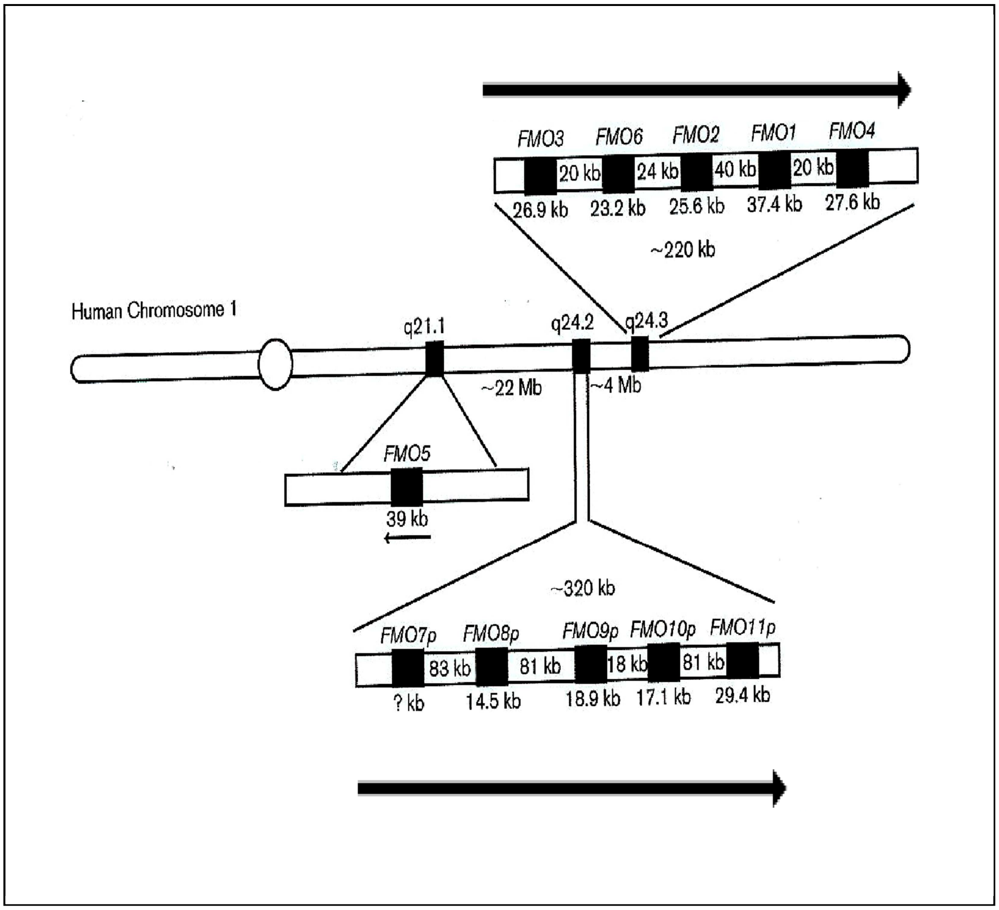


Figure 1. Localizations of the human FMO gene clusters on chromosome 1. Taken from (Hernandez *et al*. 2004).

Functional FMO genes contain eight coding exons and either one or two 5' non-coding exons (Hernandez *et al*. 2004). FMOs 1, 2, 3, 4 and 5 have 51 to 57% amino-acid sequence identity, whereas FMO6 is 72% identical to FMO3. FMOs of other mammalian species have >80% sequence identity to their human orthologues (Phillips *et al*. 2007).

1.3 Mechanism of action of FMOs

FMO belongs to a class of monooxygenases capable of generating a stable C4a peroxyflavin intermediate (Massey 1994). In the first step of the catalytic cycle, FAD undergoes a two-electron reduction by NADPH (Figure 2). The reduced flavin reacts rapidly with molecular oxygen to form the peroxyflavin intermediate, and it is in this state that FMO may exist predominantly in the cell, waiting for a suitable nucleophilic substrate with which to react (Krueger and Williams 2005). This nucleophilic attack by the substrate on FADOOH results in one atom of molecular oxygen being transferred to the substrate and another to form water (this step is important because it makes the substrate soluble, which in turn facilitates excretion). The rate-limiting steps in the catalytic cycle are believed to be the breakdown of the FADOH pseudobase or the release of NADP+ (Krueger and Williams 2005; Cashman and Zhang 2006).

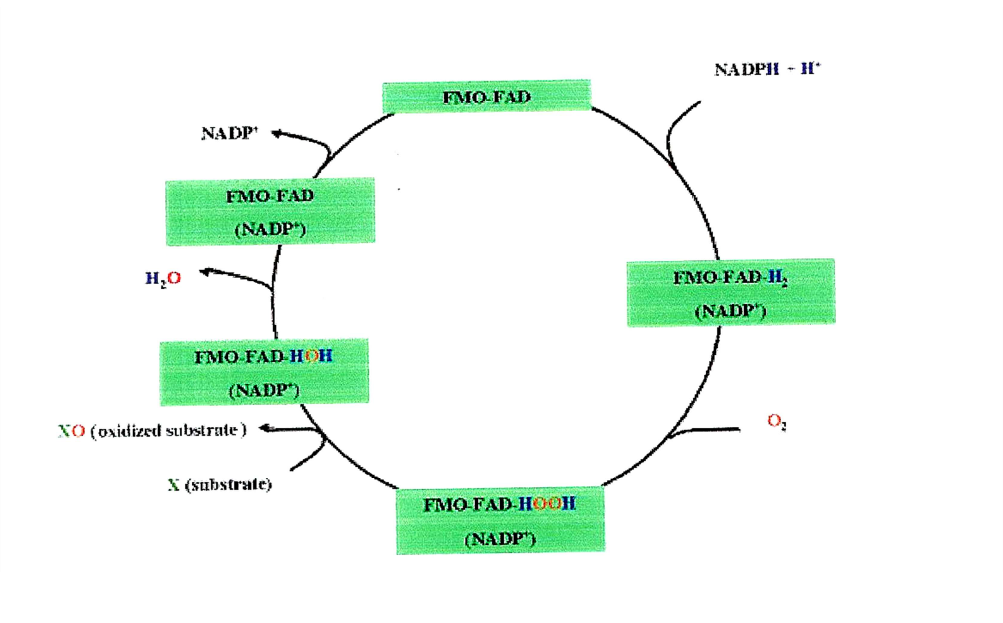


Figure 2. FMO catalytic cycle, adapted from (Kathamart and Stresser 2000).

1.4 Structure of FMO

Because human FMO proteins have not been crystallized yet, active site models have been postulated based on substrate profiles (Nagata *et al.* 1990). Information about the structure can be predicted from related proteins with crystal structures (Krueger and Williams 2005). In particular residues 12-23 of FMOs IGGGPGGLAAAR, are 83% identical to a glutathione reductase sequence, IGGGSGGLASAR, which is a ß-sheet and turn region that forms the `floor` of the FAD-binding site (Mittl and Schultz 1994). A model for the active site of several FMOs (rabbit and pig FMO1 and rabbit FMO2) has been proposed (Polyzos 2003). In this model (Figure 3), the substrate-binding channel leads to two compartments: one larger less accessible pocket (A) that binds large, planar aromatic residues, and another smaller pocket (C) that can accommodate short n-alkyl chains or a p-tolyl moity, (Cashman 1995; Fisher and Rettie 1997; Polyzos 2003).

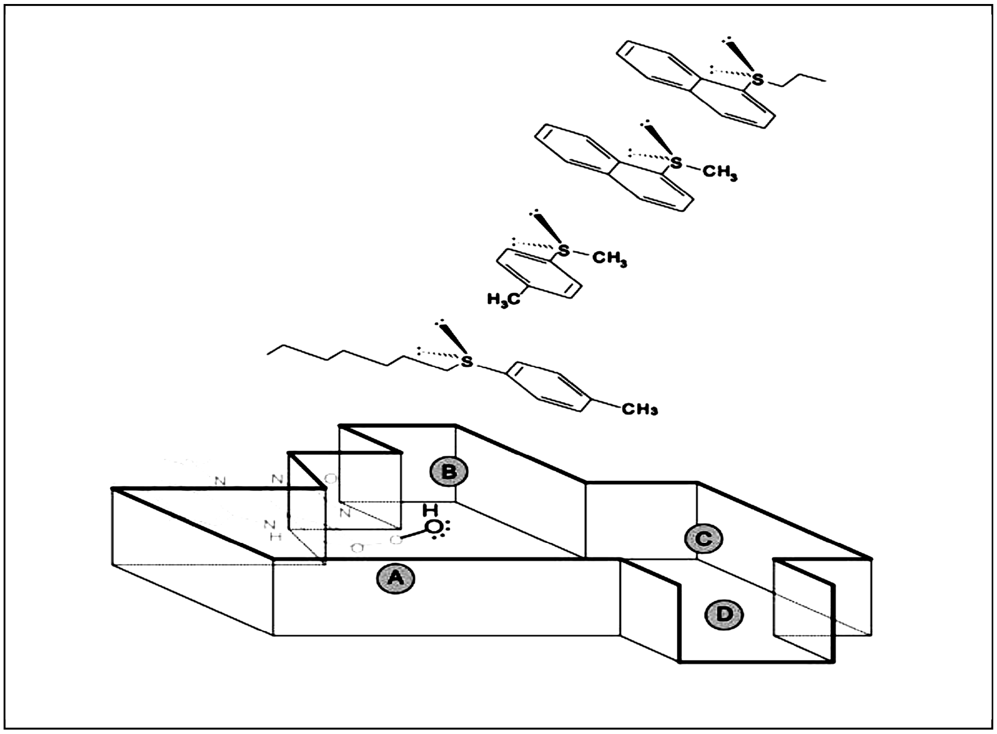


Figure 3. Model of the active site of rabbit and pig FMO1 and rabbit FMO2 proposed by Cashman (Cashman 1995) and Fisher (Fisher and Rettie 1997), showing the orientation substrates would adopt in the active site. There are two principal substrate-binding pockets, A and C, adjacent to the flavin cofactor, B and the binding channel D is open to the surface of the enzyme. (Picture taken from (Polyzos 2003).

The substrate-binding channel is clearly distinct among forms of FMO studied so far and provides enzymes with stereo- and substrate-selectivity. Larger substrates readily oxidized by pig, guinea pig and rabbit FMO1 are excluded from the active site of human FMO1, indicating a large variation among orthologues (Nagata *et al*. 1990; Lomri *et al*. 1993; Polyzos 2003). FMOs are generally 533-535 amino-acid residues long (Figure 4).

All FMOs contain two GXGXXG motifs, the FAD-binding domain at residues 9-14 and the NADPH-binding domain at residues 191-196 (Stehr *et al*. 1998). The FAD-binding site is contained within a fingerprint sequence which predicts a βαβ secondary structure, the "Rossman fold" (Wieranga *et al*. 1985; Ziegler 1993). A hydrophobic motif, characteristic of FMOs, is present at residue 330 (Krueger and Williams 2005). All mammalian FMOs studied to date have a strong membrane association, which is revealed by their poor solubility in aqueous solvent and the highly intractable nature of the purified proteins (Guan *et al*. 1991).

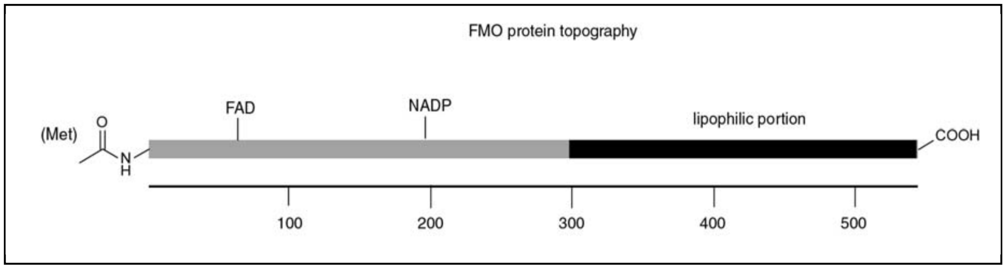


Figure 4. Linear representation of the ~535 amino acid polypeptide of rabbit FMOs 1, 2 and 3. In the N-terminal region there is the FAD-binding domain followed by an NAD(P)H-binding domain, and a lipophilic portion in the C-terminal region (Taken from (Polyzos 2003).

1.4.1 The three-dimensional structure of yeast and bacterial FMOs

The three dimensional structure has been determined of a yeast and bacterial FMO (Eswaramoorthy *et al*. 2006; Alfieri *et al*. 2008). FMO from the yeast Saccharomyces cerevisiae has 18-23% amino-acid sequence identity to mammalian FMOs (Eswaramoorthy *et al*. 2006). Unlike mammalian FMO, the yeast enzyme does not accept xenobiotics as substrates, but may be involved in maintaining cellular reducing power, probably through its action on cysteamine (Suh *et al*. 1996). Bacterial FMO from Methylophaga sp. Strain SK1 has a two-domain structure, with well defined FAD and NADP+ binding domains. Both domains are connected through a flexible hinge (Choi *et al*. 2003).

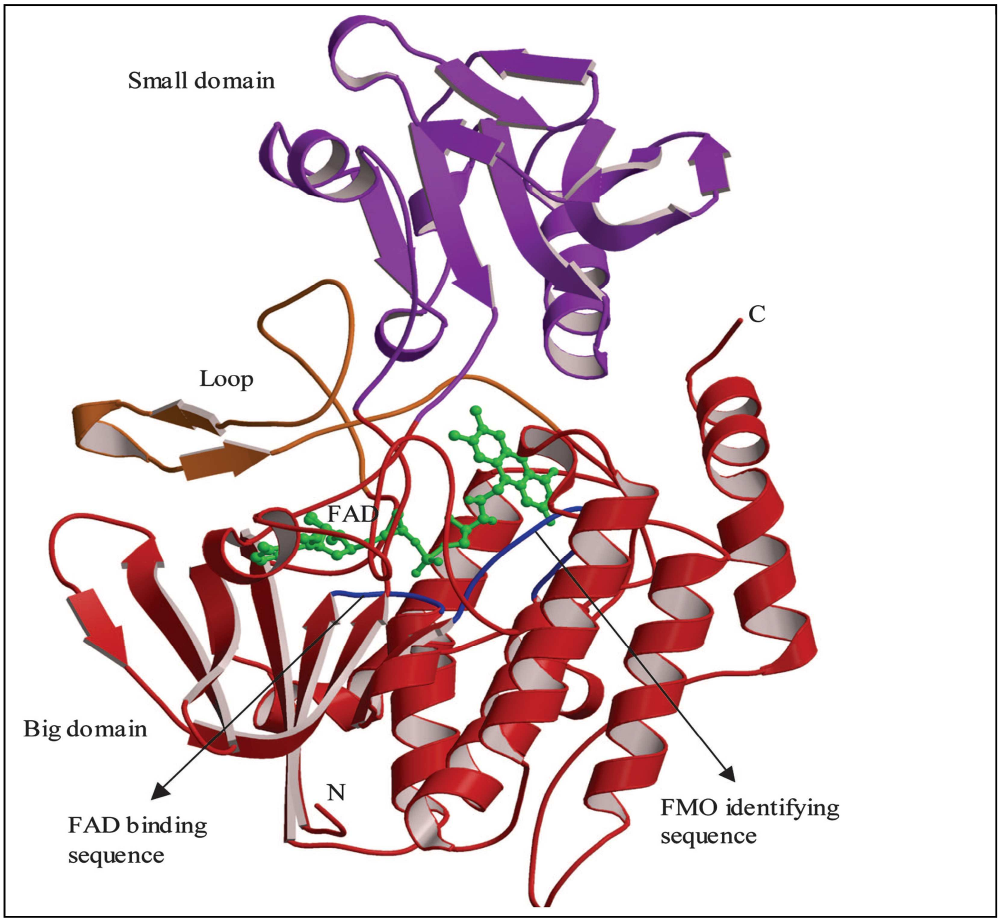


Figure 5. Ribbon representation of FMO of Yeast. FAD binds to the large domain. The strand-turn-helix motifs and the interlinking loop are labelled. Diagram from (Eswaramoorthy *et al*. 2006).

1.5 Function of FMOs

Little information exists about the functions of FMOs, other than in xenobiotic metabolism. However, it was observed that some individuals with defective FMO3 displayed abnormal metabolism of biogenic amines, which are involved in the control of blood pressure, suggesting that FMO3 may play a role in the regulation of blood pressure (Duescher *et al*. 1994).

Ziegler and Poulsen in the 1970s, hypothesized that FMO may play a role in the synthesis of disulfide bonds of protein, through cysteamine oxidation (Poulsen and Ziegler 1977; Poulsen and Ziegler 1979).

1.6 *FMO* gene expression

Each *FMO* gene exhibits a distinct developmental and tissue-specific pattern of expression in all species including humans (Dolphin *et al*. 1992; Hernandez *et al*. 2004).

1.6.1 *FMO1* gene

Human *FMO1* has 10 exons. Exons 0 and 1 are non-coding. The protein-coding sequence is located within exons 2 to 9, and the translation initiation codon is located in exon 2.

In humans *FMO1* is not expressed in the adult liver. This is in contrast to other mammals, such as rabbit, pig, rat and mouse, in which FMO1 constitutes the major form of the enzyme in the liver of adult animals. This species-dependent difference in expression may be because of the presence, upstream of exon 0, of a sequence which acts as a powerful transcriptional repressor (Shephard *et al*. 2007).

The kidney is the main site of expression of *FMO1* in adult human (Dolphin 1991; Phillips *et al*. 1995; Dolphin *et al*. 1996; Yeung *et al*. 2000). The gene is also expressed in the stomach, small intestine and in various endocrine tissues, including adrenal cortex and medulla, pancreas, thymus, thyroid and testis (Hernandez *et al*. 2004).

1.6.2 *FMO3* gene

The *FMO3* gene comprises 9 exons, of which the first is non-coding (Dolphin *et al.* 1997). In humans, the expression of *FMO3* is switched on after birth (Koukouritaki 2002), and it becomes the major isoform in adult liver (Lomri *et al.* 1992; Phillips *et al.* 1995; Dolphin *et al.* 1996; Zhang and Cashman 2006). FMO3 mRNA has also been detected in adrenal medulla and cortex, pancreas, thyroid, gut and brain, lung and kidney (Hernandez *et al.* 2004; Phillips *et al.* 2007).

1.6.3 *FMO4* gene

The *FMO4* gene contains ten exons, of which eight are coding (Hernandez *et al.* 2004). FMO4 mRNA in contrast to other FMO mRNAs contains sequences from all ten exons. *FMO4* encodes a protein that is 558 amino-acids long. This is longer than other FMOs, which contain between 532 and 535 residues (Phillips *et al*. 1995). It was suggested by Phillips *et al* (2007), that this additional block of residues are a result of a single point mutation in the termination codon of ancestral *FMO4* (Phillips *et al*. 2007). *FMO4* is mainly expressed in the liver and kidney (Dolphin *et al*. 1996; Hernandez *et al*. 2004; Zhang and Cashman 2006).

1.6.4 *FMO5* gene

FMO5 is the most ubiquitously expressed member of the *FMO* gene family and the protein that it encodes does not catalyze the oxygenation of common FMO substrates except for short-chain amines (Overby *et al.* 1997), which suggests that it may be more involved in endogenous metabolism than in the detoxification of xenobiotics. *FMO5* is expressed in many foetal and adult tissues (Benedetti and Keith 2007; Phillips *et al.* 2007).

The abundance of mRNAs encoding *FMO3* and *FMO5* are similar in the adult liver (Zhang and Cashman 2006). The contribution of FMO5 to hepatic metabolism has not been clearly established, although *FMO5* represents approximately 46% or more of the total *FMO* transcripts in adult human liver (Zhang and Cashman 2006). The different amounts of FMO mRNAs in different tissues suggests that the promoter regions of these *FMOs* contain regulatory elements responsive to transcription factors specifically present and active in these tissues.

1.6.5 FMO2 gene

The *FMO2* gene comprises of 9 exons, the first of which is non-coding (Hernandez *et al.* 2004). FMO2 is the predominant isoform in human foetal and adult lung (Cashman and Zhang 2006; Hines 2006), but, in most humans, functional *FMO2* is not expressed because of a nonsense mutation: g.23238C>T at position 472, that changed a glutamine-codon to a premature stop-codon (Dolphin *et al.* 1998). Thus, in the majority of individuals, although FMO2 mRNA is present in the lungs, no catalytically functional enzyme is expressed. In contrast, functional *FMO2* is expressed in the lungs of nonhuman primates (Cashman and Zhang 2006). *FMO2* is also expressed in skeletal muscle, kidney, prostate gland and blood vessels (Hernandez *et al.* 2004).

1.7 FMO substrate specificity

FMOs catalyze the oxygenation of a diverse group of compounds, which contain a soft nucleophile such as nitrogen, selenium, sulfur or phosphorus. The size and shape of the substrates are important factors that limit access to the catalytic site of FMOs, and these two factors are primarily responsible for differences in the substrate specificities of the FMO isoforms.

Substrates include primary, secondary (Cashman *et al*. 1999) and tertiary amines (Wu and Ichikawa 1995; Narimatsu *et al*. 1996) as well as thiourea, thioacetamide, methimazole, cysteamine and thiobenzamide (Ziegler 1991). FMOs differ in substrate specificity and range.

1.7.1 Endogenous and xenobiotic substrates

As previously mentioned, an absolute structural requirement for an FMO substrate is a soft-nucleophilic heteroatom, typically a nitrogen or sulfur, with fewer examples of selenium- or phosphorus-containing substrates (Krueger and Williams 2005). Uncharged or singly positively charged compounds are the best substrates. Although compounds with a single negative charge can be substrates, the charge must be located a certain distance from the oxygenation site (Krueger and Williams 2005). Compounds with more than a single charge are almost universally excluded from the FADOOH site (Krueger and Williams 2005). These charge restrictions for access to the substrate channel leading to FADOOH exclude many potential nucleophilic endogenous substrates from FMO-dependent oxygenation (Krueger and Williams 2005).. Many substrates for FMOs can be chemically synthesized in the body (endogenous), or are a result of conjugation of a xenobiotic with an endogenous nucleophile such as cysteine, (Krueger and Williams 2005). Examples of nitrogen-containing endogenous substrates of FMOs are the biogenic amines tyramine and phenethylamine, which are N-oxygenated by FMO to the N-hydroxy metabolite, followed by a rapid second oxygenation to the trans-oximes (Lin and Cashman 1997a; Lin and Cashman 1997b). Since the oximes appear to have little pharmacological activity, this *N*-oxygenation pathway is a mechanism for inactivation (Krueger and Williams 2005).

Another example of an endogenous compound is trimethylamine, precursors of which are present in the diet. Trimethylamine is primarily formed *in vivo* from the breakdown of choline and is *N*-oxygenated by FMO3 (Lambert *et al.* 2001; Cashman *et al.* 2004; Krueger and Williams 2005). Trimethylamine is extremely odorous (Mitchell 2005), whereas the *N*-oxide has little or no odour. Failure to efficiently *N*-oxygenate trimethylamine leads to a genetic disorder known as trimethylaminuria or fish-odour syndrome (Dolphin *et al*. 1997; Mitchell *et al.* 1997; Mitchell and Smith 2001)

Sulfur-containing endogenous substrates for FMO have also been documented (Poulsen 1981). Examples of sulfur-containing endogenous substrates of FMO are sulfhydryls, such as cysteamine, which is *S*-oxygenated to the disulfide (cystamine) (Poulsen and Ziegler 1977; Poulsen and Ziegler 1979). Cysteamine has been used therapeutically to slow down the progression of Huntington`s disease (Berman and Greenamyre 2006) and as a radioprotector (Biaglow *et al*. 1984). Cysteamine is the only approved treatment for paediatric cystinosis (Belldina *et al.* 2003). Cysteamine hydrochloride can regulate various hormones: Cysteamine hydrochloride supressess circulating growth hormone levels in male pigs (McElwain *et al*. 1999) and is a potent inhibitor of growth hormone-inhibiting hormone (Somatostatin) (Brown *et al.* 1983) FMO oxygenation of cysteamine may help control the overall thiol/disulfide redox state of the cell. By this indirect way, FMO may control the level of H_2O_2 in the cell and the expression of genes regulated by H_2O_2, sulfhydryl/disulfide ratios and the general redox state of the cell (Khomenko *et al*. 2004).

1.7.1.1 FMO1 protein

FMO1 has the broadest substrate range, including a wide range of pesticides and endogenous compounds and therapeutic drugs. A detailed explanation of how each of these drugs is metabolized by FMO1 is beyond the scope of this thesis. For more information see (Krueger and Williams 2005). However, one drug was selected, tamoxifen, because of potential conflict between detoxification and bioactivation, as well as its widespread use in the treatment breast cancer. FMO1 mediates the *N*-oxygenation of tamoxifen, which is considered as a detoxification process. So in tissues expressing FMO1, there would be expected to be fewer tamoxifen-related DNA adducts (a DNA adduct is a stretch of DNA which binds cancer-causing chemicals) (Hemminki *et al*. 1996). In contrast, the drug can be bioactivated by CYP-mediated hydroxylation. So the combined effects of FMO1 and CYP genetic variants may influence the balance between the efficacy and harmful side effects of tamoxifen (Parte and Kupfer 2005).

1.7.1.2 FMO3 protein

FMO3 catalyzes the oxygenation of nicotine, tertiary amines, such as trimethylamine and many therapeutic drugs (Cashman 2000).

Reduced activity of FMO3 can enhance the anticancer effect of nonsteroidal anti-inflammatory drugs (especially cyclooxygenase 2 inhibitors such as sulindac). Sulindac is a prodrug that contains a racemic *S*-oxide, which is reduced to the sulfide by gut bacteria (Duggan *et al*. 1977; Etienne *et al*. 2003). The active sulfide (i.e., sulindac sulfide) is *S*-oxygenated to the sulfoxide and then to the sulfone by FMO3 (Hamman *et al*. 2000). Polymorphisms of *FMO3* that decrease the functional activity of the enzyme may decrease retrooxygenation of sulindac sulfide to inactive sulfoxide and thus, increase the efficacy of sulindac, by increasing the amount of sulindac sulfide to the patient`s circulation (Etienne *et al*. 2003*;* Hamman *et al*. 2000). FMO3, Like FMO1, catalyzes the metabolism of thiacetazone, through *S*-oxygenation.

1.7.1.3 FMO4 protein

The expression of a stable form of FMO4 in heterologous systems has been proved difficult, consequently, little is known about its substrate specificity (Itagaki *et al*. 1996).

1.7.1.4 FMO5 protein

In contrast to other FMOs, FMO5 has weak catalytic activity towards compounds such as methimazole (Rettie *et al*. 1994; Cherrington 1998; (Overby *et al*. 1995). FMO5 catalyzes the *N*-oxygenation of short-chain aliphatic primary amines like *N*-octylamine. FMO5 has been reported to catalyze the oxygenation of thioethers, such as the anti-rheumatic drug easonarimod (Ohmi *et al*. 2003), which are not substrates for other FMOs. Consequently, FMO5 is regarded as an atypical FMO.

1.7.1.5 FMO2 protein

FMO2 differs from other FMOs, for it can catalyze the *N*-oxygenation of some primary alkylamines to their oximes (Lin and Cashman 1997a; Lin and Cashman 1997b), but is unable to catalyze the oxygenation of certain tertiary amines, such as imipramine and chlorpromazine (Ohmiya and Mehendale 1982; Williams *et al*. 1984). FMO2 has a more restricted substrate range, with a preference for substrates with longer side chains, suggesting that the active site of FMO2 is located further from the surface of the enzyme than that of other FMOs and is accessed by a relatively narrow channel (Williams *et al*. 1984). FMO2 is also less sensitive to inactivation by anionic detergents and elevated temperature (Guan *et al*. 1991).

Substrates include thioether-containing organophosphate pesticides, such as phorate and disulfoton (Henderson *et al*. 2004). The products of the FMO2-catalyzed reaction are less toxic than the parent compounds (Neal and Halpert 1982) and therefore, the enzyme has a protective role.

The role of FMO2 in drug metabolism is still unclear. However, the antipsychotic drugs prochloperazine and trifluoperazine are substrates of rabbit FMO2 (Lomri *et al*. 1993).

Substrates of particular interest are thiacetazone and ethionamide (Henderson *et al*. 2008; Francois *et al*. 2009), the second-line antitubercular drugs (Brown 1992; Peloquin 1993; Alahari *et al*. 2007) which are activated by the mycobacterial enzyme EtaA (Vanelli *et al*. 2002). Pulmonary tuberculosis is a widespread disease, especially in Africa: according to the World Health Organization 31% of all the new cases in 2006 were in Africa: *(http://www.who.int/mediacentre /factsheets /fs104/en/).*

In contrast to most other substrates of FMOs, which are metabolized into less toxic substances, the thiourea in these drugs is bioactivated by FMO2-mediated *S*-oxygenation to toxic sulfenic and/or sulfinic acid metabolites (Figure 6). Sulfenic acid is capable of undergoing redox cycling following conjugation with glutathione, resulting in oxidative stress due to depletion of reduced glutathione and NADPH (Smith and Crespi 2002; Henderson *et al*. 2004) (Figure 6). This process is believed to be an important mechanism of thiourea toxicity in the lung (Henderson *et al*. 2004).

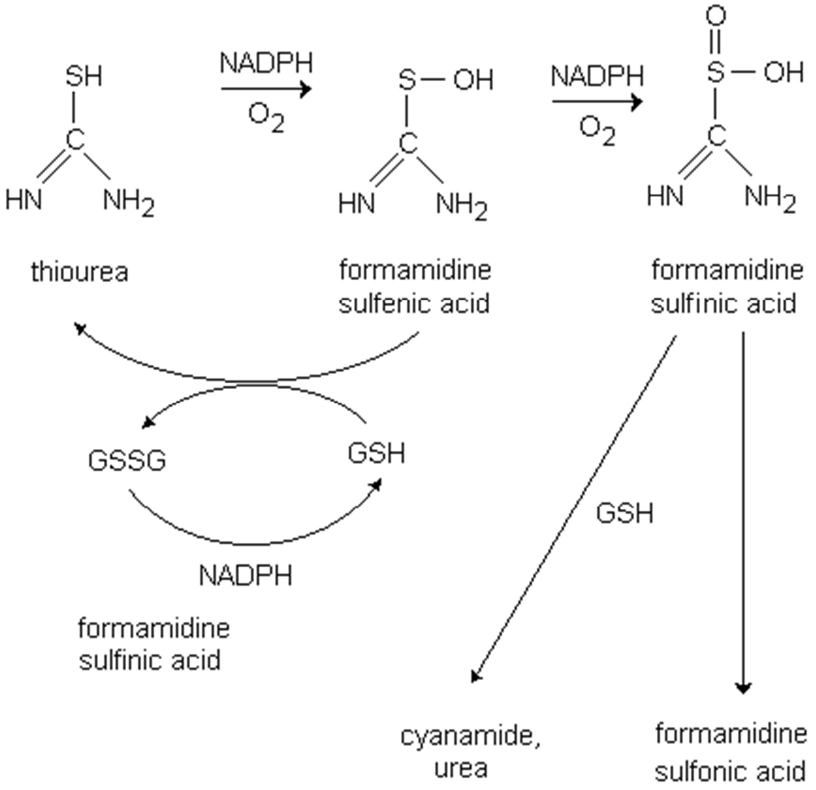

SH
C
HN
NH2
NADPH
O2
S— OH
NADPH
O2
O
S— OH
thiourea
formamidine sulfenic acid
formamidine sulfinic acid
GSSG
GSH
NADPH
formamidine sulfinic acid
GSH
cyanamide, urea
formamidine sulfonic acid

Figure 6. Pathway of FMO2-mediated *S*-oxygenation of thiourea. Diagram from the international Chemical Assessment document 49 (http://www.inchem.org/documents/cicad).

1.8 Association of FMO with human disease

A global gene expression investigation of patients diagnosed with atrial fibrillation documented a significant increase in the expression of FMO1 mRNA (Kim *et al*. 2003). Studies of the spinal cord of patients with amyotrophic lateral sclerosis indicated that FMO1 mRNA was under-represented (Malaspina *et al*. 2001), also, the expression of certain FMO enzymes in some neoplastic tissues has been observed (Rebhan *et al*. 1997). There is suggestive evidence that FMO deficiency may be associated with sideroblastic anaemia (Barber *et al*. 2000). FMO3 deficiency causes trimethylaminuria (TMAU), also known as fish-odour syndrome (Dolphin *et al*. 1997; Mitchell and Smith 2001; Mitchell 2005), which is manifested clinically by a body odour similar to that of rotten fish and is accompanied by various psychological abnormalities (Ayesh *et al.* 1993; Mitchell and Smith 2001). It is classified as a rare but ancient disorder. As the name implies, the disorder is associated with an excess of extremely odorous trimethylamine in the urine, sweat and breath of the patient (Mitchell and Smith 2003). Trimethylamine is normally metabolized to the *N*-oxide, which has no offensive odour, by FMO3 in the liver (Sardas *et al*. 1996; Mitchell *et al*. 1997). The type of mutation in the *FMO3* gene defines the severity of the disease (Park *et al*. 2002; Hernandez *et al*. 2003; Zhang *et al*. 2003).

Diagnosis is by determination of urinary ratios of trimethylamine *N*-oxide to trimethlylamine following a challenge dose of choline. The severity of the disease can be predicted to be greater with a lower ratio (Mitchell and Smith 2001).

1.9 *FMO2* variants

Two major *FMO2* alleles are present in humans: *FMO2*1*, which encodes a full-length active protein and *FMO2*2A*, which encodes a truncated inactive protein (Dolphin *et al.* 1998; Whetstine *et al*. 2000).

A comparative analysis of the cDNA for FMO2 of human with that of FMO2 of rabbit (Lawton *et al*. 1990), guinea pig (Nikbakht *et al*. 1992) and Rhesus macaque (Yueh *et al*. 1997), revealed that the former encoded a truncated protein which lacks 64 amino-acid residues from its carboxy terminus (Dolphin *et al*. 1998). This was caused by a nonsense mutation (See section 1.6.5 page 11).

Heterologous expression of the cDNA revealed that the truncated polypeptide was catalytically inactive (Dolphin *et al*. 1998). The nonsense mutation that gave rise to the truncated polypeptide, is not present in the *FMO2* gene of closely related primates, including gorilla and chimpanzee (Dolphin *et al.* 1998), and must therefore have arisen in the human lineage after the divergence of the *Homo* and *Pan* clades which took place some 5-6 million years ago (Brunet *et al*. 2002) and has subsequently spread to attain a frequency of close to 100% in present day human populations (Asians and Europeans) (Whetstine *et al*. 2000; Furnes *et al*. 2003; Phillips *et al*. 2007).

The protein product of *FMO2*1* allele is 97% identical to FMO2 of Rhesus macaque and 85% identical to FMO2 of rabbit and mouse (Dolphin *et al*. 1998).

FMO2 is usually inactive in humans, it is found in an active form only in Africans and individuals of recent African origin (Whetstine *et al*. 2000).

Interestingly 27% of Africans (Dolphin *et al*. 1998; Whetstine *et al*. 2000), as well as 5% of Hispanics (Krueger *et al*. 2004; Hickey 2007) possess at least one allele coding for the full-length protein.

FMO2 gene expression has been demonstrated to be regulated by sex hormones in experimental animals (Lee *et al*. 1993; Dolphin *et al*. 1998), and putative glucocorticoid-responsive elements have been identified in the 5'-flanking region of the rabbit *FMO2* gene (Wyatt *et.al* 1996).

The similarities in the pattern and size of expression of FMO2 mRNA in humans and other mammalian species indicate that the human *FMO2* gene has suffered no mutations that affect either the expression of the gene or the processing or stability of the corresponding mRNA (Dolphin *et al*. 1998). So the absence of FMO2 in human lung might be explained by the lack of 64 residues from its carboxyl terminus. The truncated protein might be unable to fold correctly, and thus would be detected by cellular surveillance systems, such as the ubiquitin pathway, and rapidly degraded (Dolphin *et al*. 1998). The lung plays an important role in the metabolism of inhaled foreign chemicals, including drugs, environmental toxicants and carcinogens (Zhang *et al*. 2006b). Human FMO2 thus represents an unusual case of a gene that has become non-functional in humans, but not in other primates.

Furnes *et al*. (2003) have identified additional variants in *FMO2* in African-Americans. Krueger *et al*. (2005) investigated the effect of the occurrence of four common polymorphic variants of *FMO2* [g.7700-7702 insGAC (71Ddup) in exon 3, g.10951delG (V113fsX) in exon 4, g.13732C→T (S195L) in exon 5 and g.22060T→G (N413K) in exon 8)] identified in African-Americans (50 individuals) by Furnes *et al*. (2003) and in Hispanics by Krueger *et al*. (2005). Of these, only the g.22060T>G (N431K) variant had full catalytic activity. The allelic frequency of these variants was lower in Hispanics than in African-Americans. Krueger *et al*. (2005) demonstrated through inferred haplotype determination that these variants are located on the major, g.23238T allele which encodes a truncated, non-functional protein and, thus, would not significantly impact the activity of the full-length functional enzyme produced by the g.23238C allele (Krueger *et al*. 2005). Veeramah *et al*. (2008) conducted a study on the frequency of the functional allele *FMO2*1* in 24 groups from different regions of Africa, as well as from Yemen and Turkey, and found that the distribution of the allele is relatively homogeneous across sub-Saharan Africa with approximately one-third of individuals possessing at least one *FMO2*1* allele, though in some populations the incidence of these individuals approached 50%. Coding-region variants of *FMO*s are shown in Table 1.

Table 1. Coding-region variants for *FMOs 1, 2, 3, 4* and *5*

Gene	Variant	exon	Amino acid change	Functional consequence	Reference
FMO1	**g.9614C>G**	**3**	**H97Q**	**no effect**	**(Furnes *et al*. 2003; Furnes and Schlenk 2004)**
FMO1	**g.22739G>A**	**6**	**R223Q**	**n.d.**	**dbSNP126**
FMO1	**g.22818C>T**	**6**	**T249T**	**-**	**(Furnes *et al*. 2003), dbSNP126**
FMO1	**g.23970A>G**	**7**	**I303V**	**no effect**	**(Furnes *et al*. 2003; Furnes and Schlenk 2004); dbSNP126**
FMO1	**g.23971T>C**	**7**	**I303T**	**no effect**	**(Furnes *et al*. 2003; Furnes and Schlenk 2004); dbSNP126**
FMO1	**g.25061A>G**	**8**	**V396V**	**-**	**(Furnes *et al*. 2003), dbSNP126**
FMO1	**g.27258A>G**	**9**	**P467P**	**-**	**dbSNP126**
FMO1	**g.27362C>T**	**9**	**R502X**	**substrate-dependent decrease**	**(Furnes *et al*. 2003; Furnes and Schlenk 2004)**

FMO2	g.107A>G	2	D36G	n.d.	(Furnes *et al*. 2003)
FMO2	g.7661G>A	3	V59I	no effect	(Dolphin *et al*. 1998)
FMO2	g.7695T>A	3	F69Y	n.d.	dbSNP126
FMO2	g.7700_7702dupGAC	3	D71dup	loss of function	(Furnes *et al*. 2003; Krueger *et al.* 2005)
FMO2	g.7731T>C	3	F81S	n.d.	dbSNP126
FMO2	g.10951delG	4	V113fsX	loss of function	(Furnes *et al*. 2003; Krueger *et al.* 2005)
FMO2	g.13693T>C	5	F182S	n.d.	(Furnes *et al*. 2003)
FMO2	g.13732C>T	5	S195L	loss of function	(Furnes *et al*. 2003; Krueger *et al*. 2005)
FMO2	g.13733A>G	5	S195S	-	(Furnes *et al*. 2003)
FMO2	g.18237G>A	6	R238Q	n.d.	(Furnes *et al*. 2003)
FMO2	g.18269C>T	6	R249X	n.d. (but likely loss of function)	dbSNP126
FMO2	g.19679A>G	7	E314G	n.d.	dbSNP126
FMO2	g.19839G>A	7	A367A	-	(Furnes *et al*. 2003)
FMO2	g.19898_19899ins TCAAGCTC	7	R387RfsX5	n.d. (but likely loss of function)	(Furnes *et al*. 2003)
FMO2	g.19910G>C	7	R391T	n.d.	(Furnes *et al*. 2003)
FMO2	g.22027G>A	8	E402E	-	(Furnes *et al*. 2003)
FMO2	g.22060T>G	8	N413K	no effect	(Furnes *et al*. 2003; Krueger *et al*. 2005)
FMO2	g.23238C>T	9	Q472X	loss of function	(Dolphin *et al*. 1998; Whetstine *et al*. 2000)
FMO2	g.23300A>G	9	K492K	-	(Furnes *et al*. 2003)
FMO2	g.23405_23406insT	9	F528FfsX32	n.d.	dbSNP126
FMO2	g.23412_23413insT	9	C530LfsX30	n.d.	(Whetstine *et al*. 2000)

FMO3	g.72G>T	2	E24D	limited	(Koukouritaki *et al*. 2007)
FMO3	g.11177C>A	3	N61K*	reduced or abolished	(Koukouritaki *et al*. 2007)
FMO3	g.15089G>C	4	D132H	substrate-dependent decrease	(Furnes *et al*. 2003; Lattard *et al*. 2003)
FMO3	g.15167G>A	4	E158K	moderate, substrate-dependent decrease	(Dolphin *et al*. 1997; Brunelle *et al*. 1997; Treacy *et al*. 1998; Akerman *et al*. 1999a; Zschocke *et al*. 1999; Furnes *et al*. 2003)
FMO3	g.15475G>T	5	G180V	no effect	(Dolphin *et al*. 2000)
FMO3	g.15550C>T	5	R205C	moderate decrease	(Fujieda *et al*. 2003)
FMO3	g.18281G>A	6	V257M	no effect or limited substrate -dependent decrease	(Treacy *et al*. 1998; Dolphin *et al*. 2000; Furnes *et al*. 2003)
FMO3	g.18290A>G	6	M260V	n.d	(Shimizu *et al*. 2006)
FMO3	g.21350T>C	7	V277A	n.d	(Cashman and Zhang 2002)
FMO3	g.21443A>G	7	E308G	moderate substrate-dependent decrease	(Treacy *et al*. 1998; Akerman *et al*. 1999b; Zschocke *et al*. 1999)
FMO3	g.21599T>C	7	L360P	increased activity	(Furnes *et al*. 2003; Lattard *et al*. 2003)
FMO3	g.21604G>C	7	E362Q	n.d.	(Cashman and Zhang 2002; Furnes *et al*. 2003)
FMO3	g.23613G>T	8	K416N	limited	(Koukouritaki *et al*. 2007)
FMO3	g.24642G>A	9	I486M	n.d.	(Cashman and Zhang 2002)
FMO3	g.24691G>C	9	G503R	n.d.	(Furnes *et al*. 2003)

FMO4	**g.110T>C**	**2**	**I37T**	**n.d.**	**(Furnes *et al.* 2003)**
FMO4	**g.14601C>T**	**7**	**F281F**	**n.d.**	**dbSNP126**
FMO4	**g.14680A>T**	**7**	**T308S**	**n.d.**	**dbSNP126**
FMO4	**g.14724T>C**	**7**	**D322D**	**n.d.**	**(Furnes *et al.* 2003)**
FMO4	**g.14726T>C**	**7**	**V323A**	**n.d.**	**(Furnes *et al.* 2003)**
FMO4	**g.14770G>C**	**7**	**E339Q**	**n.d.**	**(Furnes *et al.* 2003)**

FMO5	**g.23716G>A**	**7**	**P337P**	**n.d.**	**dbSNP126**
FMO5	**g. 23806A>G**	**7**	**A367A**	**n.d.**	**(Furnes *et al.* 2003)**
FMO5	**g.37917C>T**	**9**	**P457L**	**n.d.**	**(Furnes *et al.* 2003)**
FMO5	**g.38059G>T**	**9**	**R506S**	**n.d.**	**dbSNP126**
FMO5	**g.23716G>A**	**7**	**P337P**	**n.d.**	**dbSNP126**

Mutation nomenclature follows that recommended by the Human Genome Organization (http://www.hgvs.org/mutnomen/). *Likely to be causative for TMAuria. Data in table is from Phillips *et al.* (2007) (Phillips *et al*. 2007). There are 28 additional rare mutations for FMO3, see (Phillips *et al*. 2007). SNP, single-nucleotide polymorphism.

1.10 Natural selection

Natural selection works to either decrease or increase the frequency of alleles that have a detrimental or favourable effect on an individual`s fitness (Campbell and Tishkoff 2008). There are several types of selection including positive selection. In positive selection, an advantageous allele can increase in frequency, together with linked variants (i.e., genetic hitchhiking), and decrease the genetic variation in a given population (i.e., a selective sweep) (Harris and Meyer 2006; Nielsen *et al.* 2007; Sabeti *et al.* 2007). This will result in strong linkage disequilibrium (LD) (the non-random association between two loci not necessarily on the same gene) (Sabeti *et al.* 2002; Voight *et al*. 2006; Tishkoff 2007b). The size of the sequence being subjected to selection depends on the local recombination rate and on the power of selection. In time, recombination will result in the accumulation of new mutations, which will in turn breakdown the extent of LD (Campbell and Tishkoff 2008). Other forms of selection include purifying and balancing selection. Purifying selection acts to keep deleterious mutations at low frequency in a population due to their negative effect on fitness. Therefore, weak purifying selection also displays a pattern of an excess of low-frequency mutations (Campbell and Tishkoff 2008). Balancing selection leads to an increase in intermediate-frequency alleles, it usually occurs when heterozygotes have an advantage over homozygotes (Campbell and Tishkoff 2008). In this form of selection, multiple alleles (different versions of a gene) are selected at frequencies higher than expected under neutral evolution.

The evolutionary patterns of sequences subjected to selection are not easily detectable, since demographic processes can mimic some of these patterns. An example of a demographic process is population size changes. Population expansion, for example, results in an increase in low-frequency variants, which is similar to the pattern observed due to positive or purifying selection (Campbell and Tishkoff 2008), whereas a population bottleneck (drastic decrease in population size either due to death or failure to reproduce) results in an increase in genetic drift (the change in frequency of an allele due to random sampling across generations). Genetic drift may cause the disappearance of an allele and therefore reduces genetic variation by decreasing the frequency of rare variants. This results in a pattern similar to the one caused by balancing selection and population structure (Nielsen *et al*. 2007).

It is believed that the migration of some modern humans out of Africa into diverse environments has resulted in non-African populations having more recently selected loci than African populations (See section 1.11) (Akey *et al*. 2004; Storz *et al*. 2004; Hawks *et al*. 2007; Williamson *et al*. 2007). It should be noted that, although the exposure to a novel environment may result in adaptive pressures in the exposed populations, the lower rate of selection detected in Africa could be due to the studies being restricted to a limited number of African populations and that African-Americans have often been used to represent African populations. The admixture of African-Americans with Europeans (Carlson *et al*. 2005; Williamson *et al*. 2007), will result in lower power to detect selection, since population structure can mimic balancing selection (Oleksyk *et al*. 2010). There are many candidate genes for natural selection, for example, the HbS mutation in the ß-globin gene, which causes sickle cell anaemia in homozygous individuals. Individuals who are heterozygous for the sickle cell trait are protected against malarial infection and have higher reproductive fitness (Kwiatkowski 2005). Other examples include the *G6PD* gene (Tishkoff *et al*. 2001), the Duffy gene on chromosome 1 (Hamblin and Di Rienzo 2000), genes involved in dietary adaptations (e.g., genes involved in lactase persistence (Hollox *et al*. 2001; Itan *et al*. 2009) and bitter-taste receptors (Kim *et al*. 2005). Other candidate genes include ones involved in antibody-antigen reactions or encoding drug-metabolizing enzymes like FMO2 (Stahl and Bishop 2000; Nielsen *et al*. 2005). It is of interest that evidence has been reported for balancing selection of *FMO3* (Allerston *et al*. 2007).

1.11 Human origins and evolution

Anthropogeny is the study of the evolution and the origins of the human race. The word "human" originates from the Latin humanus, which is the adjective of homo. A detailed description of human origins is beyond the scope of this thesis. The interested reader can refer to (Stringer and Andrews 2005).

Gaining insights into the evolutionary history of modern humans is important for a better understanding of how human genetic variation relates to susceptibility to disease and to differences in response to various pharmacological agents (mainly therapeutic drugs).

Africa has more human genetic diversity in both nuclear and mitochondrial genomes (Tishkoff and Verrelli 2003; Garrigan *et al*. 2007; Jakobsson *et al*. 2008) than observed anywhere else on Earth (Tishkoff *et al*. 2009). This variation is due to Africa`s complex population history, in that humans existed for a long time and in bigger population sizes, making Africa the ideal geographic location for studying human genetic variation (Atkinson *et al*. 2008; Campbell and Tishkoff 2008).

There has been a paucity of large-scale autosomal studies of populations that are ethnically diverse, especially in Africa (Conrad and Hurles 2007). The importance of these studies lies in the identification of population-specific variants, as well as variants playing roles in susceptibility to complex diseases in people of recent African ancestry (Tishkoff and Williams 2002; Reed and Tishkoff 2006; Sirugo *et al*. 2008).

In an attempt to estimate the frequency of common SNPs shared among major population groups in the United States of America, Guthery *et al*. (2007) resequenced 154 chromosomes from African-American, European, Asian and Latino/Hispanic populations for 3873 genes. This study revealed higher genetic diversity of individuals with recent African origin (African Americans). African Americans had the lowest frequency of common SNPs (36%) and the highest frequency of rare SNPs (64%), of which 44% were population-specific (Guthery *et al*. 2007).

Haplotype and nucleotide variation studies indicated that the geographical and historical separation of African populations predates the out-of-Africa migration (Garrigan *et al*. 2004; Plagnol and Wall 2006; Yotova *et al* .2007).

This observed pattern of genetic variation in Africa reflects the region`s demographic history, such as admixture and population size, as well as natural selection, recombination and mutation (see Section 1.10 and Campbell and Tishkoff 2008). The decrease in genetic diversity moving out and away from Africa (Ramachandran *et al*. 2005; Liu *et al*. 2006) is believed to be the result of a bottleneck (Forster and Matsumina 2005; Macaulay *et al*. 2005) associated with the migration of modern humans out of Africa.

There are two models for the migration of humans from Africa, the out-of-Africa model (OOA), developed by Chris Stringer and Peter Andrews (Stringer and Andrews 1988) and the multiregional model (Wolpoff *et al*. 2000), which remains a possibility due to the presence of archaic admixture in modern humans with Neanderthales (Green *et al*. 2010; Reich *et al*. 2010).

The OOA model suggests that modern *Homo sapiens* originated in Africa ~200,000 years ago and then spread across the rest of the world within the past ~100,000 years (Tishkoff *et al.* 2003). This has been supported by fossil remains from Ethiopia, dated to ~150-190 kya (White *et al*. 2003; McDougall *et al.* 2005).

The OOA model has favoured two routes, one being migration into the Levant, via the Nile Valley and North Africa, followed by subsequent migration into both Europe and Asia (Campbell and Tishkoff 2008). The second, earlier route, was across the Bab-el-Mandeb strait at the mouth of the Red Sea, travelling along the Indian ocean into Southeast Asia and Australia, ~50,000, and then into Europe, ~30,000 years ago (Forster and Matsumina 2005; Reed and Tishkoff 2006) (Figure 7). Archaeological data, as well as results from both mitochondrial and Y-chromosome DNA studies suggest that anatomically modern humans originated in eastern Africa and migrated to southern Africa (Wood e*t al*. 2005), although a southern African origin cannot be ruled out (Campbell and Tishkoff 2008; Henn *et al.* 2011). The multiregional model (Wolpoff *et al.* 2000) is thought to have involved an initial dispersal of *Homo erectus* from Africa ~1,000,000 years ago, and expanded into Eurasia.

The multiregional model has been contradicted by the observed shared patterns of genetic diversity among non-African populations, e.g, at the *CD4* locus, and the presence of old mtDNA haplotypes in Eurasia, South Asia and Australia, which indicates that all non-African populations derive from a single ancestral gene pool (Tishkoff *et al.* 1996) and this, in turn, supports the out-of-Africa model (Quintana-Murci *et al.* 1999; Macaulay *et al.* 2005; Thangaraj *et al.* 2005). But despite this, the multiregional (multiple dispersal) model remains a possibility because of the presence of archaic admixture in modern humans (see above) (Green *et al*. 2010; Reich *et al*. 2010).

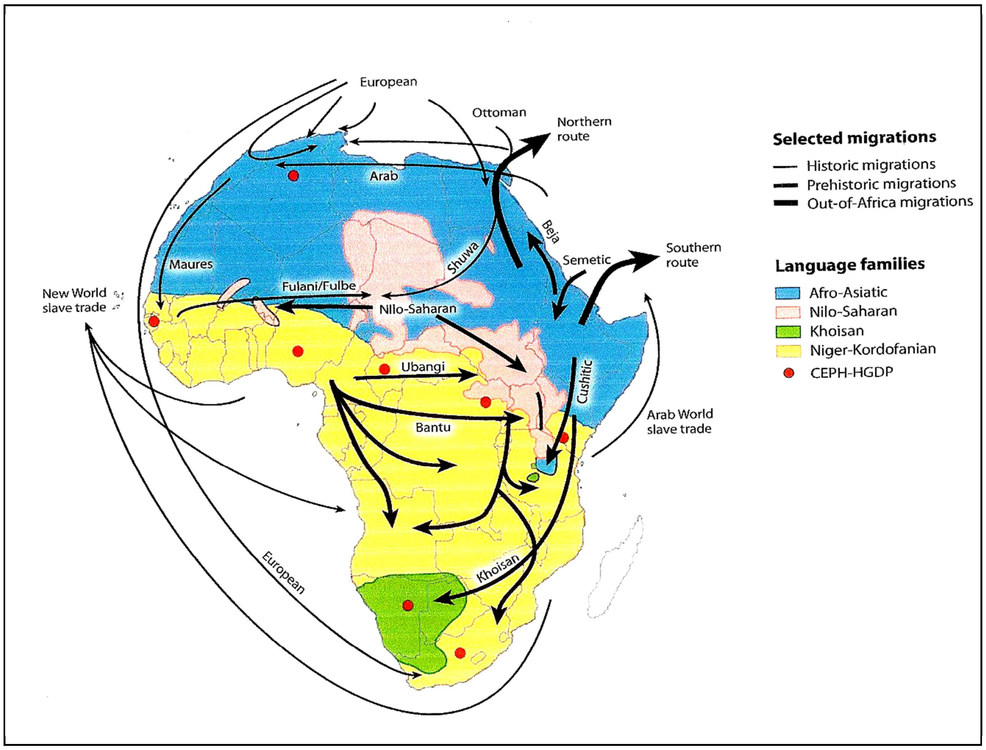


Figure 7. Map of African Language Families and hypothesized migration out and within Africa adapted from (Reed and Tishkoff 2006). African languages are divided into four families: Afro-Asiatic, Nilo-Saharan, Khoisan and Niger-Kordofanian (Campbell and Tishkoff 2008). Red dots indicate the locations from where the African samples included in the Centre d`Etude du Polymorphisme Humain (CEPH) Human Genome Diversity Panel (CEPH-HGDP) originated (Reed and Tishkoff 2006).

Recently, interest has arisen in examining human genetic variability in Africa. One of the studies involved examined 400 INDELs and 800 short tandem-repeat polymorphisms (STRPs) by genotyping more than 2000 ethnically diverse African individuals. The study revealed high levels of admixture between most African populations (at least 13) (Tishkoff *et al*. 2009), with the only exception being Bantu speakers (Niger-Kordofanian), who are more genetically homogeneous compared with other African populations. This is believed to be caused by the resent expansion of Bantu from a common origin in Cameroon/Nigeria (see Section 1.13.5).

1.12 Geographical features of sub-Saharan Africa

Sub-Saharan Africa consists of the area that lies south of the present day Sahara Desert (Quintana-Murci *et al*. 1999; Von Cramon-Taubadel and Lycett 2008).

The Sahel stretches across northern Africa from the Red Sea to the Atlantic Ocean. It is the transitional zone between the Sahara desert in the north and the wetter rainforests and savannas of Equatorial Africa in the south (Figure 7).

The horn of Africa (sometimes called Somali Peninsula) is located in the north-east of Africa along the southern side of the Gulf of Aden. The Horn of Africa and areas of Sudan, although being part of the Arab world (see above), are also geographically part of sub-Saharan Africa (Figure 8).

The population of sub-Saharan Africa was ~800 million in 2008 (Group 2010).

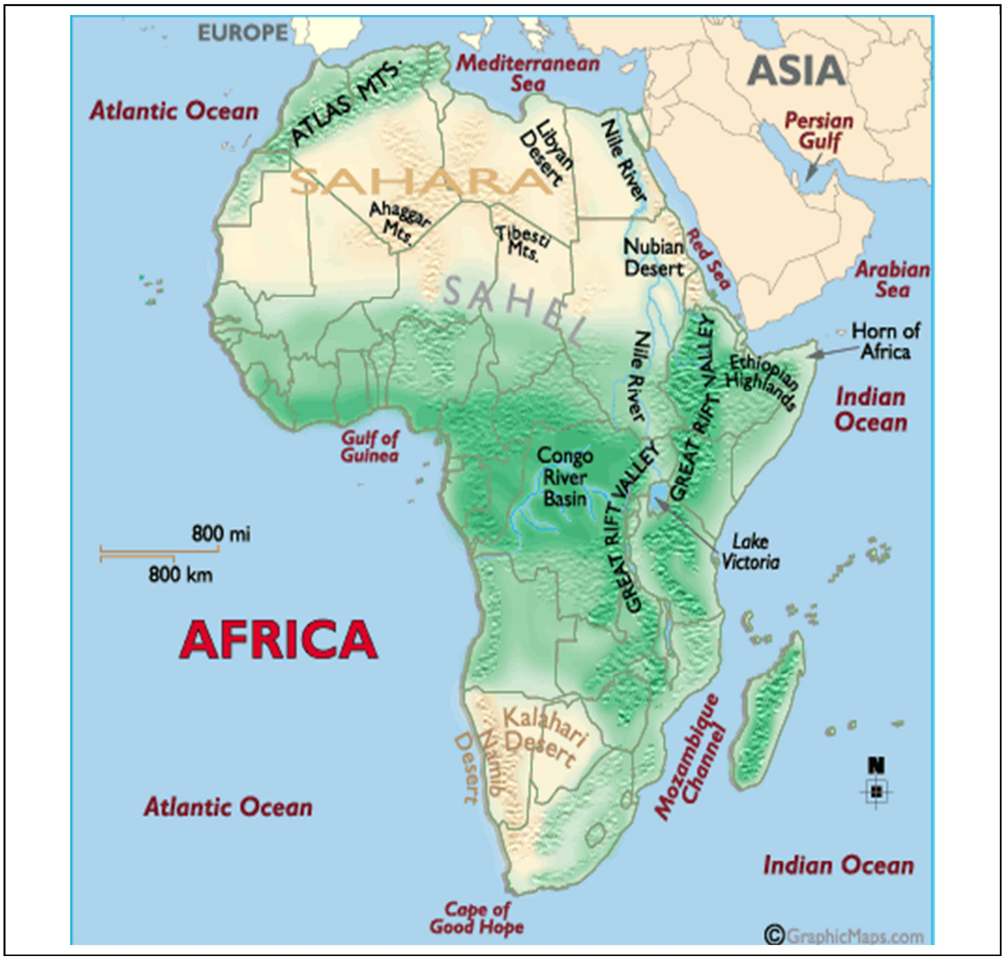


Figure 8. Geographic map of Africa: Map illustrates the Saharan Desert, the Sahel, the Sahara Desert, the Horn of Africa in the north and northeast, respectively, and the Kalahari basin in the southwest, light and dark green are tropical savannas and rainforests of Equatorial Africa, Map taken from (GraphicMaps.com).

1.13 Sub-Saharan African populations

1.13.1 West-African populations

1.13.1.1 Ethnic groups from Cameroon

1.13.1.1.1 The Fulbe

The ethnic name ‘Fulbe’ is the plural of ‘Pullo’, and the autonym for the Fulbe people as an ethnic group (also known as the Fulani) (Barreteau *et al.* 1984). The Fulbe are Muslim pastoralists who started migrating into Cameroon from present-day Nigeria in the beginning of the 13th century (Neba 1999). The Fulbe included in this study are from Mayo Darle, Adamawa, in northern Cameroon.

There are estimated to be 165,000 individuals (http://www.mandaras .info/Fulbe.html). Fulfide (Fulbe), is their language. It is a west-Atlantic language (Niger-Kordofanian).

1.13.1.1.2 Shuwa Arabs

Shuwa Arabs are semi-nomadic, agricultural people who are descendants of Arab tribes originating from present-day Saudi Arabia *(http://encyclopedia2.the freedictionary. com / Shuwa+Arab)*. They are Muslims. The Shuwa Arabs used in this study reside in Sadigo, a populated region in Cameroon. They speak Arabic, which is an Afro-Asiatic Language (www.Ethnologue.com).

1.13.1.1.3 Mambila

The Mambila live on both sides of the Nigeria/Cameroon border, the majority living on the Mambila plateau in Nigeria. A smaller number (c.12,000) are found in Cameroon, especially on the Tikar Plain (Zeitlyn 1994). The Mambila included in this study are from the Somie village. The villagers are self-sufficient in food, since they grow several crops including rice, maize, bananas, sweet potato and tobacco. They have also grown coffee as a cash crop since the 1960s (Rehfish 1960). Most people in the village are members of either the Catholic or Protestant church. They speak Mambile, which is a member of the Niger-Kordofanian language family (www.Ethnologue.com).

1.13.1.2 Ethnic groups from Ghana

1.13.1.2.1 The Asante

The Asante, also called the Ashanti, are a major ethnic group of the Ashanti region in Ghana. They are Twi-speaking Akan people. Twi is a member of the Niger-Kordofanian language family. The Ashanti settled in central present-day Ghana. The Ashanti make up 19% of the Ghanaian population (~7 million). They continue to be influential, although their power has fluctuated since Ghana's independence. The majority are Christians (Protestant), with a Muslim minority (Crentsil 2007).

1.13.1.2.2 The Bulsa

The Bulsa are an ethnic group of about 174,000 people practicing subsistence agriculture and cattle holding (Schott 1987). They are from Sandema, the capital of the Bulsa district. They speak Buli, a Niger-Kordofanian language. Their main religion is Christianity, but there are also Muslims. The Bulsa are a group quite distinct from all their neighbours (Schott 1977).

1.13.1.3 Ethnic groups from Senegal

1.13.1.3.1 Manjak

The Manjak people are an ethnic group in Guinea-Bissau. The Manjak also have a strong community in Senegal from which the samples used in this project were collected. They number 201,000 in Senegal *(http://www.joshuaproject.net/ people-profile.php).* The Manjak language belongs to the Niger-Kordofanian language family (www.Ethnologue.com).

1.13.2 Central-east African populations

1.13.2.1 Ethnic groups from Tanzania

1.13.2.1.1 The Chagga

The Chagga are the third largest ethnic group in Tanzania (2,000,000 individuals) (Yakan 1999). They are Bantu-speaking people who migrated into the hills of Kilimanjaro from northern Cameroon (Greenberg 1949).

The Chagga reside on the slopes of both Mount Kilimanjaro and Mount Meru, as well as in the Moshi area. The Chaggas included in this study live around Mount Kilimanjaro. The climate, as well as successful agricultural methods, are reasons for the relative wealth of these people. Another factor that gave them an economic advantage over other African population groups is that they were one of the first tribes to convert to Christianity (Yakan 1999). The Chagga mainly work in agriculture, banana being their main food. But their best known export is Arabica coffee, resulting in coffee being a primary cash crop (Ehret 2002).

1.13.3 South-east African populations

1.13.3.1 Ethnic groups from Malawi

1.13.3.1.1 Malawi

Malawi is a south-east African country with various environments. The people of Malawi belong to various Central Bantu groups. The Chewa and Nyanja group, collectively known as Malawi (or Maravi), make up 58% of Malawi`s population. The main language of the Chewa is Nyanja, but the two most spoken languages are Chichewa (also called Chewa) and Chitumbuka. These languages are members of the Niger-Kordofanian language family (www.ethnologue.com).

The Maravi were the first Bantu people to move into present-day Malawi. The Chewa live on the Lilongwe plains (Reader 1999), while the Nyanja are located in Chikwawa in southern Malawi after migrating there from central Zaire. The main religion is Christianity, although the traditions of ancestral worship still play an important role for most Nyanja (Ntara 1973). Samples from Malawi in this study are from the Chewa.

1.13.3.2 Ethnic groups from Mozambique

1.13.3.2.1 Sena

The Sena are a Mozambican ethnic group, which is prominent in the Zambezi valley. There are 1,500,000 individuals. They speak Sena, which is member of the Niger-Kordofanian family (www.ethnologue.com). The people of Mozambique have managed to retain their small-scale agricultural culture, despite the influence of European colonizers and Islamic coastal traders (*http://www.theseedcompany. org /project/sena-nt*).

1.13.4 East-African populations

1.13.4.1 Ethnic groups from Ethiopia

1.13.4.1.1 The Afar

The Afar are an ethnic group who live in the Afar region of Ethiopia, Eritrea and Djibouti (Weeks 1984). They make up 2% of the total Ethiopian population (1,300,000), ~10% of whom are urban inhabitants, according to the most recent census (2007) http://www.csa.gov.et/pdf/Cen2007_firstdraft.pdf. The Afar language is part of the Cushitic branch of the Afro-Asiatic language family. They are mainly nomadic-pastoralists and are Muslims.

1.13.4.1.2 The Amhara

The Amhara are mostly farmers inhabiting the north central highlands of Ethiopia. They speak the Amhara language which has semetic origin represented by sharing alphabet and words with both Hebrew and Arabic. Their language is related to other Semitic languages like Gurage (Jenkins 1997).

The Amhara are thought to originate from modern-day Yemen. This explains their skin colour (olive skin), and their southern Arab features. Their ancestors migrated from the Arabian peninsula into Eritrea and Ethiopia (Jenkins 1997). The majority of the Amhara are Christian, but around 18% are Muslims.

1.13.4.1.3 The Anuak

The Anuak are a river people. They live in villages scattered along the river banks of western Ethiopia, in the Gambela region and the rivers of southeastern Sudan. They number between 100,000-150,000 http://en.wikipedia.org/wiki /Anuak_people. The Anuak are ethnically, historically, religiously and linguistically different from most other Ethiopians, and so is their indigenous land. The Anuak, unlike other Nilotic people in the region (Nilotic people or Nilotes, refers to some ethnic groups in southern Sudan, Uganda, northern Tanzania and Kenya, who speak Nilotic languages, a large sub-group of the Nilo-Saharan languages), who raise cattle, are herdsmen and farmers.

The Gambela region is tropical and hot with well-watered soil from the highland rivers. This difference in geography has separated Ethiopians into “lowlanders”, such as the Anuak, and “highlanders,” such as the Amharas (see above) and the Oromos (see below). The Anuak and other lowlanders have been the victims of discrimination due to their dark skin colour, for they are considered to be black Africans, as opposed to most other highland Ethiopians, who are of lighter colour (Collins 1971).

A recent study based on cluster analysis that looked at a combined sample of Amhara and Oromo found that they share 62% of their genome with Caucasians (Askenazi Jews, Norwegians and Armenians), 24% with other sub-Saharan Africans (Bantus), 8% with Austro-Melanesians (Papua New Guineans) and 6% with East Asians (Chinese) (Wilson *et al.* 2001). The Caucasoid contribution to the Ethiopian gene pool was estimated to be >60% (Cavalli-Sforza *et al.* 1994) and occurred predominantly through males (Wilson *et al.* 2001). Conversely, the Niger-Kordofanian contribution to the Ethiopian population occurred mainly through females (Quintana-Murci *et al.* 2010).

1.13.4.1.4 The Gurage

The Gurage are an ethnic group in southwest Ethiopia in a semi-mountainous region. They make up 2.5% of the total Ethiopian population (1,900,000, of which 50% are urban dwellers), according to the 2007 national census http://en.wikipedia.org/wiki/Gurage_people.

The Gurage included in this study are from Indibir and Buta Jira. The Gurage live a life based on agriculture, involving a complex system of crop rotation and transplanting, their main crops include staple crops such as rice and maize (Nida 2005).

The Gurage languages, more than other Ethiopian semetic languages, have been influenced by Cushitic languages, mainly the Oromo, which is the most populous language with 35 million speakers (Shack 1966). More than 50% of the Gurage are Christian (Orthodox), while 40% are Muslims (Nida 2005).

1.13.4.1.5 The Nuer

The Nuer are tribes which live in southern Sudan and western Ethiopia. They are mainly pastoral people. The Nuer have gradually migrated east into Ethiopia. The Nuer included in this study were from the lowlands of Gambela and Akobo. It is named after the Akobo River, which flows westwards then north into the Bro river, defining Ethiopia`s border with Sudan (Evans-Pritchard 1940). They speak Nuer, which is a Nilo-Saharan language (www.ethnologue.com).

1.13.4.1.6 The Oromo

The Oromo are an ethnic group found in Ethiopia, northern Kenya and in parts of Somalia *(http://en.wikipedia.org/wiki/ Oromo_people)*. The have migrated into Ethiopia from the land of Cush, present-day Nubia. They make up ~35% of population groups (30 million), according to the 2007 census *(http://en.wikipedia.org/wiki/ Oromo_people)*. Therefore they make up the single largest ethnic group in Ethiopia. They are mostly agriculturalists. The language of the Oromo is Oromiffa (Afaan Oromo), a Cushitic language belonging to the Afro-Asiatic language family (Shinn *et al*. 2004). Their religions include traditional, Christianity and Islam.

(Bates 1979) contends that the Oromo "were a very ancient race, the indigenous stock, perhaps, on which most other peoples in this part of eastern Africa have been grafted". (See Figure 9) for the locations from where all sub-Saharan African samples were collected).

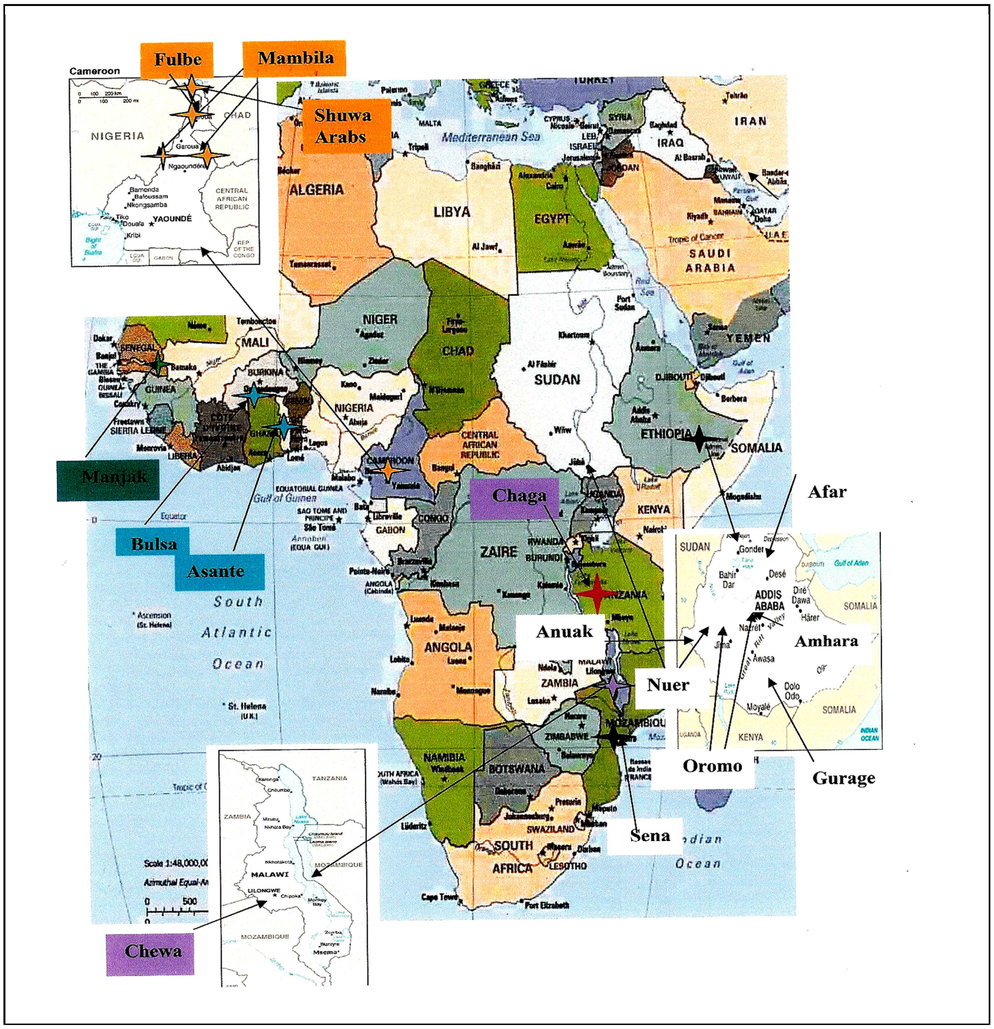


Figure 9. Map displaying the geographical locations from where African samples were collected

1.13.5 The Bantu Expansion

The Bantu are a diverse group made up of 300-600 ethnic groups. Bantu also refers to a large language category. The Bantu language belongs to the Niger-Kordofanian language family. An example of a Bantu language is Swahili. Although it is not a populous language (only 5-10 million speakers), its importance lies in that it has been taken up as a second language by millions of Africans.

At present, the majority of Bantus live in southern, central and east Africa. This is a result of the Bantu expansion. The Bantu expansion, or migration, was the long series of migrations of the ancestors of Bantu speakers (proto-Bantu, or pre-Bantu) from what is present-day Nigeria and Cameroon (Clark 1984; Adler 2007). The exact date of the first migration is unclear, but it is believed to have begun after agriculture had already been introduced, which would indicate a date around 3000-2500 BC (http://en.wikipedia.org/wiki/ Bantu_expansion).

This is thought to have resulted in an increase in population size and, therefore, increased pressure to expand and cultivate new land. Bantu speakers had developed skills in iron-working, which were useful in the manufacture of weapons and tools for cutting trees for shifting “slash-and-burn” cereal agriculture. These factors lead to an early expansion within west Africa (phases I and II), followed by eastwards and southwards migrations beyond west Africa into present day Angola around 1500 BC (phase III) (Figure 10) (Vansina 1995). The Bantu expansion has been considered to be one of the most important human migrations within the past thousand years due to the impact of Bantu admixture with indigenous populations on genetic variation of modern Africans (Wood *et al*. 2005; Tishkoff *et al*. 2007a).

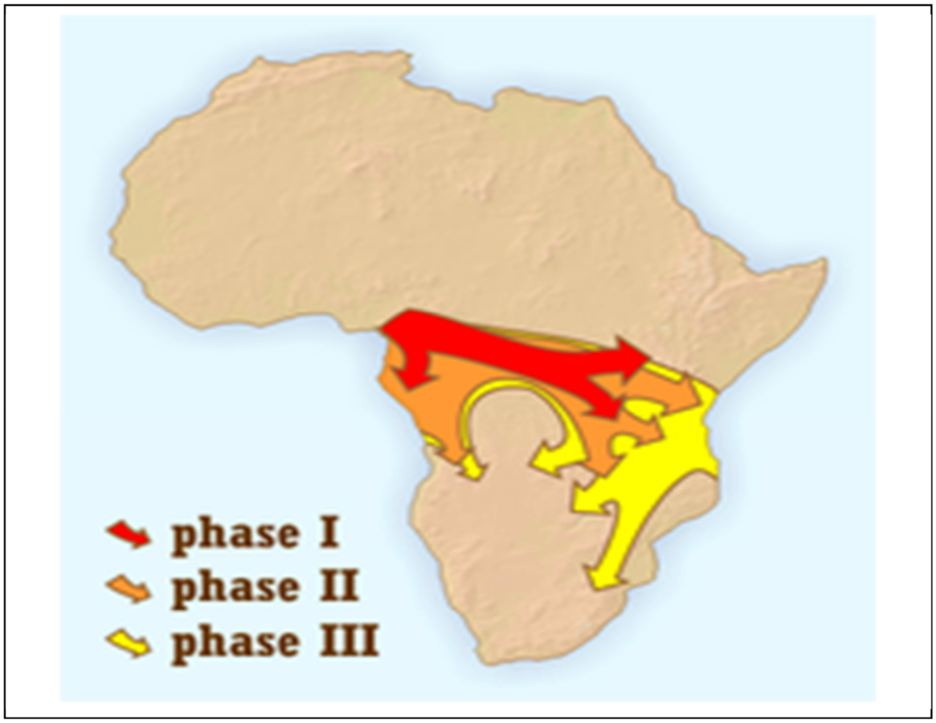


Figure 10. The three phases of the Bantu expansion from West Africa.

1.14 Aims

Because of the implications of expressing functional FMO2, especially in regard to drug efficacy and public safety, it was important to assess the distribution of the *FMO2*1* and *FMO2*2A* alleles in sub-Saharan Africa, as well as the evolutionary relationship among mutations. The role of functional FMO2 in the metabolism of xenobiotics, and thus its role in mediating interactions between humans and their chemical environment, makes FMO2 a likely target for natural selection.

One aim was to determine whether potentially deleterious SNPs occur on a full-length functional *FMO2*1* or on a truncated non-functional *FMO2*2A* background.

Another aim was to estimate the frequency of functional FMO2, and to investigate this, samples were initially genotyped for six SNPs reported by the

National Centre for Biotechnology information (NCBI) to be present at high frequency in Africa, and then resequenced to examine whether other nonsynonymous SNPs occur on a full-length background. This will help provide insights into the evolutionary history and future of the *FMO2* gene.

A further aim was to determine the evolutionary relationships of the mutations, i.e., the order in which they occurred, this was performed using network, whereas GeneTree was used to estimate the date to the most recent common ancestor (T_{MRCA}) and to date some of the mutatons.

FMO2 being an enzyme involved in the metabolism of foreign chemicals, makes it a likely candidate for natural selection, therefore *FMO2* sequences were be tested for evidence of selection.

2. Methods and Materials

2.1 Selection of samples

2.1.1 Criteria for selecting samples for genotyping

Population samples were chosen from throughout Africa (especially sub-Saharan Africa) on the basis of a high frequency in the population group of the full-length allele (*FMO2*1*), the availability of fifty or more individuals from the group and its geographical location.

2.1.2 Criteria for selecting samples for resequencing

The samples were chosen according to their genotype for the truncation SNP from two ascertainment plates available in the Centre of Genetic Anthropology (TCGA) at University College London (UCL). The samples were chosen to represent the overall frequency in Africa of both g.23238C and g.23238T FMO2 alleles.

2.2 Sample collection

All sample collection procedures mentioned have been performed by individuals other than the author of this Thesis.

2.2.1 Genotyping samples

2.2.2 Resequencing samples

Genomic DNA samples were prepared from mouth swabs taken from 747 individuals from various populations grouped into east (EA), and west-Africa (WA). Signed consent was obtained from all individuals (Available on request).

2.2.2.1 West-African ascertainment plate

The 370 DNA samples included in the plate were chosen to include the regions associated with the expansion of the Bantu speaking people throughout west-Africa (see Section 1.13.5).

2.2.2.2 East-African ascertainment plate

The 377 DNA samples included in the plate were chosen by sampling a north east to south west transect across Ethiopia to capture the majority of genetic variation based on pairwise F_{ST}s from Y chromosome and mtDNA hypervariable region 1 sequence.

2.3 Genotyping

2.3.1 Adjusting the concentration of genomic DNA samples

Genomic DNA samples were electrophoresed through 0.9% agarose gels in order to estimate their concentration. The concentration was estimated by comparison of the band intensity to a standard curve of DNA concentration. Concentrations were adjusted to 1ng/µl by addition of an appropriate volume of TE buffer.

2.3.2 Design of primers for genotyping

For designing the primers, the software file builder version 3.1, provided by Applied Biosystems (ABI), was used. Initially, a target sequence has to be chosen. This sequence is preferably 600 bases long, to increase assay design possibilities, and contains the target site (SNP) in the center. This target sequence was obtained from NCBI, according to the SNP being genotyped. The target sequence was then tested for uniqueness. This was performed using the BLAST (Basic Local Alignment Search Tool) search provided on the NCBI website to detect regions with sequence similarities and repetitive elements (www.ncbi.nlm.nih.gov /BLAST).

To further test for repetitive elements, the target sequence was run through a program called Repeat Masker (www.repeatmasker.genome.washington.edu). All ambiguities in the target sequence were then masked by an N. It was important to make sure that there were no Ns in close proximity to the target site (SNP site), since this would reduce the number of available primers and probes for optimal assay selection: http://www.icmb.utexas.edu/core/DNA/Information_Sheets /Realtime%20PCR/7900ABDsnp.pdf. The resulting file was then loaded onto file builder and was submitted to Applied Biosystems (ABI).

Samples were genotyped for six common, nonsynonymous *FMO2* SNPs, which have a high frequency in Africa. Genotyping was performed using TaqMan assays designed by ABI (40X assay mix).

2.3.3 Principle of TaqMan genotyping assay

TaqMan assays use the 5` to 3` exonuclease activity of the Taq DNA polymerase. Each reaction contains a gene-specific primer and a fluorescently labelled TaqMan probe. The probe contains a 5`reporter dye and a 3`quencher dye. The 3` end is blocked to prevent extension during PCR. The probe is designed to anneal the target sequence between the forward and reverse PCR primers. While the probe is intact, the quencher suppresses the fluorescence of the reporter dye. During amplification, *Taq* DNA polymerase cleaves the probe and removes it from the target, allowing extension to continue. Scission of the probe separates the reporter dye from the quencher dye, resulting in an increase in fluorescence (Figure 11). The increased fluorescence only occurs if the target sequence is amplified and is complementary to the probe, thus preventing detection of non-specific amplification. For any given cycle, the amount of fluorescence signal (amount of product), is directly proportional to the initial copy number (http://www.diagnostickits.in/Analytical%20principle.html).

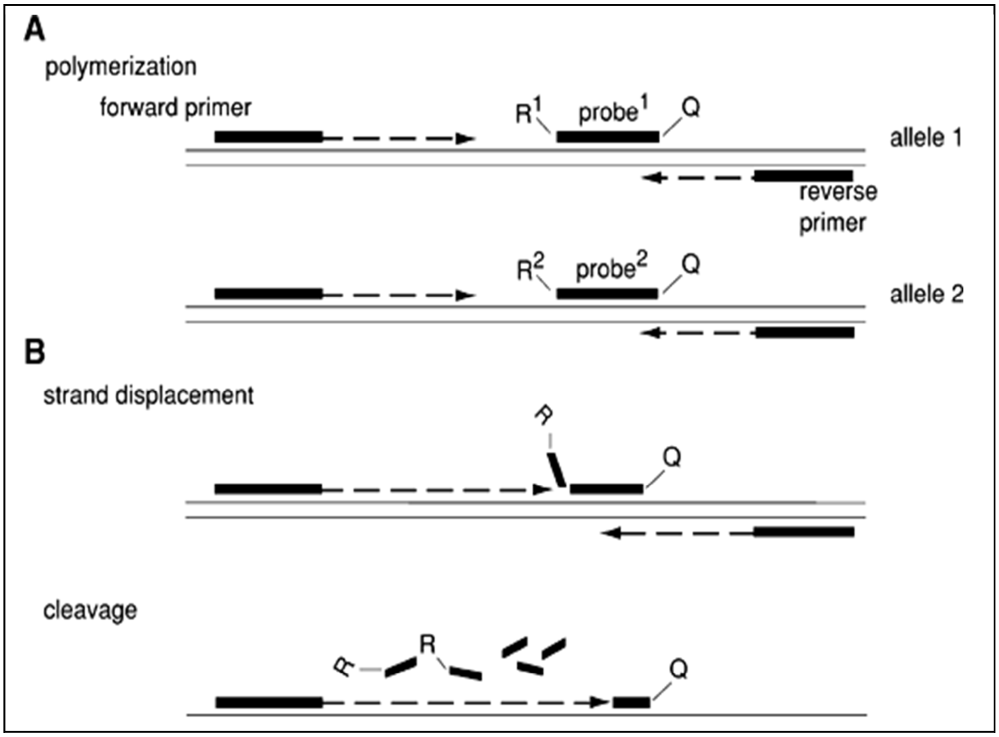


Figure 11. TaqMan assay diagram

Polymerization: A fluorescent reporter (R) dye and a quencher (Q) are attached to the 5' and 3' ends of a TaqMan probe respectively.
Strand displacement: When the probe is intact the reporter dye emission is quenched.
Cleavage: During each extension cycle the DNA polymerase cleaves the reporter dye from the probe.
Polymerization completed: Once separated from the quencher, the reporter dye emits its characteristic fluorescence (Hiller 2007).
Text adapted from Applied Biosystems.
Graphics courtesy of Molecular Probes.

2.3.4 TaqMan Protocol

DNA was amplified in 4 µl reaction volumes in 384-well microplates containing 2 µl of 1x TaqMan Genotyping Master mix (Applied Biosystems (ABI), Warrington UK), 0.74 µl of sterile water, 0.16 µl of MgCl (25 mM), 1 µl of 1ng/ µl DNA and 0.05 µl of 80x assay mix (containing probes and primers from (ABI). Thermal cycling conditions on the GeneAmp PCR System 9700 thermal cycler were as follows: an initial incubation at 95°C for 10 min, to activate the Ampli TaqGold enzyme, then 40 cycles of 92°C for 15 sec, to denature the double-stranded DNA, and 60°C for 1 min, to anneal primers and extend the DNA. An endpoint plate read of the resultant PCR product was performed using TaqMan 7900HT software (Applied Biosystems (ABI), Warrington UK).

2.3.4.1 TaqMan genotyping of g.23238 C>T

TaqMan technology (ABI), was used for the genotyping of g.23238 C>T, as well as for the other five SNPs. The primers FMO2-g.23238F, ACTCTATTTCG GACCCTGCAA, and FMO2-g.23238R, GCATTTCTGGCTCCTTCCCATT, were used to amplify a region of 511 bp containing the c.23238C>T SNP. The fluorogenic probes VIC, AGGCGATACTGATAGGAG, and FAM, AGGCGATA CTAATAGGAG, were included in the PCR to detect, respectively, the presence of a G (C on forward strand) or A (T on forward strand) at position +1414.

The probes were designed to ensure that if they annealed specifically between the forward and reverse primers, this would result in probe degradation by the 5`nuclease activity of the Taq-polymerase during PCR. Primers for the genotyping of the other five SNPs are listed in Table 2.

Table 2. Genotyping primers

SNP dbSNP Cluster#	Forward Primer	Reverse Primer	VIC	FAM
1)rs2020870	5`AGCCCACTTGCTTTGAGAGAA3`	5`TCACTCTAAAAGAACCCAGAACTTGT3	CTCCTCCAATATCTTCAG	TCCTCCAATAACTTCAG
2)rs2020860	5`GTTACCAACACCAGCAAAGAAATGT3`	5`AGATTTGAAGACCTTATAAAGTCCTAAAAACG3`	ATTATGCAGGGAGTTTG	TGCAGGAAGTTTG
3)rs28369860	5`GATGCCTTCTCTGTGTTTCTTCAAG3`	5`TCTAAAGAGTAGGAGACCGGTTAC3`	ACAACTGTCCTTAGTTGAGAA	AACTGTCCTTAGTGTGAGAA
4)rs2307492	5`CCATAGCCGCCAATACAAGCA3`	5`TAGGACCACTAACCTTACCCTTTG3`	TTTTCCCTCAGATCCAT	TTTCCCTCAAATCCAT
5)rs2020862	5`GCATCCTGGTGATTGGAATGG3`	5`GACAACTCGACTCATTCTTACGAC3`	TGAGCCTGAGTTTC	CTGAGCCTAAGTTTC

2.4 DNA resequencing

Samples (50 ng of genomic DNA) were sequenced by Macrogen (Rockville, Maryland, USA) for all eight coding exons as well as flanking intronic regions of the *FMO2* gene. The complete sequence of human *FMO2*, described in the National Institute of Environmental Health Sciences (NIEHS) SNP Database in fasta format was used as a reference and primers to amplify each exon were designed. Sense and antisense primers are described in (Table 3). Primers were selected based on the following criteria: GC contents 50% + /-3%, a length of 10-21 nucleotides. PCR was performed with Platinum polymerase (Invitrogen), with an annealing temperature of 55^{o}C as a default. PCR products were sequenced on both strands using an Applied Biosystems BigDye Termination Cycle Sequencing Kit, version 3.1. Ethanol precipitation was used to remove excess dye. An ABI PRISM 3730xl DNA analyzer (Applied Biosystems) was used to analyse the eluates (Macrogen).

Traces were viewed using Sequencher 49, demo version (http://www.gene codes.com/). Since these traces cannot be saved or exported using the demo version, sequence scanner v.1 from Applied Biosystems was used as well.

Sequence traces for each exon and flanking intronic regions were visualized using both sequencher v.4.9 and sequence scanner, and checked for the presence of SNPs (novel or previously identified), against a reference sequence obtained from GenBank (accession number AL021026). The traces were edited, and low quality regions of sequence from each individual were removed. For each exon and its flanking intronic regions, the longest stretch of high-quality sequence that was available for all individuals was taken for analysis.

Table 3. Sense and Antisense Primers for resequencing

Exon Number	Sense Primer	Antisense Primer
Exon 2	5`AAGGCAAGAAGAGAGCAGGA3`	5`GTCTTCACTATAGAACTCTTC3`
Exon 3	5`GGTAGAGACCTAGGATGCTA3`	5`GTCTCACTCTGAGACAGAGT3
Exon 4	5`CATGAGGAACCCGCTCTTTCT3`	5`GTCTTCTTGGTCAAACTACAC3`
Exon 5	5`TTCACTATGCTGCTCAGGCT3`	5`ACGTCACTCGGTACAACTAC3`
Exon 6	5`CTTAAGGAGCAGCAAACCAC3`	5`CTTTCCGTCCAGCATTAGAG3`
Exon 7	5`TGACCATAGGCACAGGCATT3`	5`TAAACGTCTATACGACTCC3`
Exon 8	5`AACCACTAGTCATGGCTGC3`	5`GAGATTGGTGTCTACATCC3`
Exon 9	5`GACACCAAGCTATGCTTT3`	5`CCGACGAAAAGTACGACACA3`

2.5 Data analysis

2.5.1 Genotyping Data

2.5.1.1 Haplotype inference

Haplotypes were inferred using Phase v.2.0 Phase is a statistical algorithm that uses a Bayesian method to reconstruct haplotypes from genotype data (Stephens *et al*. 2001). The algorithm initially identifies all unambigous haplotypes from homozygote and single-site heterozygote genotypes observed in the population sample, and then moves on to inferring all ambiguous haplotypes using a Markov chain-Monte Carlo (MCMC) algorithm (see Gilks *et al.* 1996, for an example). It does so, by repeatedly choosing an individual at random and attempts to estimate haplotypes assuming that all other haplotypes are correctly reconstructed. The software is freely available at: http://www.stat.washington.edu /stephens/phase/download.html, from where I downloaded v.2.0.

2.5.1.2 Genotypes

2.5.1.2.1 Hardy-Weinberg equilibrium (HWE)

The Hardy-Weinberg equilibrium states that both genotype and allele frequencies are in equilibrium in a random-mating population. Disturbances to this equilibrium are caused by selection, small population size, meiotic drive, random genetic drift, gene flow, non-random mating. It is calculated by the formula: $p^2+2pq+q^2 = 1$, where p is the frequency of the major allele and $q = 1-p$. An online Hardy-Weinberg calculator was used, which is accessible at *http://www.oege.org/soft ware/hwe-mr-calc.shtml* (Rodriguez *et al*. 2009). This calculator also performs the chi-square test (see below) for genotypes.

2.5.1.2.2 Pearson`s chi-square test

This test examines a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have a total probability of 1 (Plackett 1983). All Chi-square goodness of fit tests were performed using an interactive chi-square test calculator (Preacher 2001), accessible at http://www.people.ku.edu/~preacher/chisq/chisq.htm.

2.5.1.2.3 Network diagrams

The mutational relationship among the haplotypes was visualized using both the median-joining (MJ), and the reduced-median algorithms of Network.

Network is useful for the visualization of reticulation which may indicate recurrent mutation, recombination and genotyping errors. The median-joining (MJ) algorithm is used for constructing networks from recombination-free population data (Bandelt *et al.* 1999), whereas the reduced-median (RM) algorithm is preferential for simple sequences in binary format. Network generates diagrams and networks from genetic, linguistic and other data. The diagrams are represented as ball and chain diagrams, where the size of the ball is proportional to haplotype frequency (Bandelt *et al*. 1995).

All tests marked with an $^{(*)}$ were performed using DnaSP, version 5.

All test marked with (#) were performed with Arlequin, version 3.1.

2.5.2 Resequencing Data

2.5.2.1 Allele frequencies and haplotype inference*#

Allele frequencies for each SNP were determined by gene counting, on both the full-length (23238C) and truncated (23238T) backgrounds. Haplotypes were inferred using three approaches: the Phase algorithm of DnaSP, version 5 (http://www.ub.es/dnasp/), and the maximum-likelihood approach, using the EM (Excoffier *et al.* 1995) and ELB algorithms of Arlequin (Excoffier *et al.* 2003). DnaSP is a software package for Windows. It can be used to estimate several measures of DNA sequence variation between and within populations, as well as various statistical tests.

2.5.2.2 Prediction of the effect of nonsynonymous mutations on function

The PolyPhen algorithm was used to predict the possible impact of an amino-acid substitution on the structure and function of a human protein (Xi *et al.* 2004). The predictions are based on the nature of the amino-acid change and the degree of conservation across species. The algorithm performs this by using amino-acid sequence alignments and protein structure database information (such as the Uni Prot database). The predicted effect is one of the following: benign (no damage); possibly damaging; probably damaging; presume definite damage.

2.5.2.3 Network visualization of haplotypes

To gain insights into the mutational relationship between haplotypes, the median-joining (MJ) and the reduced-median algorithms of Network were used (Bandelt *et al*. 1995; Bandelt *et al*. 1999) (see Section 2.5.1.2.3).

2.5.2.4 Statistical tests

2.5.2.4.1 Fisher`s exact test*

Fisher`s exact test is used in the analysis of small sample size contingency tables. It is called an exact test because the deviation from a null hypothesis can be calculated exactly, rather than approximately (Fisher 1922). The test is useful for examining the significance of association (contingency) between two kinds of categorical classifications. Fisher`s exact test was performed using an online server (www.quantitativeskills.com/sisa/statistics/fisher.htm).

2.5.2.4.2 Four-gamete test*

The four-gamete test estimates the minimum number of recombination events (Rm) by locating pairs of segregating sites that cannot have arisen without either recurrent mutation or recombination (Hudson and Kaplan 1985).

2.5.2.4.3 Pairwise linkage disequilibrium (LD)*

Linkage disequilibrium is the non-random association of alleles at two or more loci, not necessarily on the same chromosome. Linkage describes whether alleles occur more or less frequently in a population than would be expected from random haplotype formation.

Factors which influence the level of linkage disequilibrium are genetic drift, non-random mating, population structure, rate of recombination and selection. Since the oldest measure of LD, D is of little use for measuring the strength of and comparing LD level (Lewontin 1964), another common measure of linkage disequilibrium, R^2, was used. The measure of R^2 is equal to D^2 divided by the product of the allele frequencies at the two loci. Hill and Robertson deduced that E $[R^2]=1/1=4Nc$, where c is the recombination rate in morgans between the two markers and N is the effective population size (Hill and Robertson 1968).

Significant associations between variant sites were identified by using both Fisher`s exact test and chi square test (χ_2) with a Bonferroni correction for multiple comparisons.

Z_{nS} (Kelly 1997), which is the average of R^2 over all pairwise comparisons (Hill and Robertson 1968), Z_a (Rozas *et al*. 2001), the average of R^2 (Hill and Robertson 1968) over all pairwise comparisons between adjacent polymorphic sites; ZZ (Rozas *et al*. 2001) = Z_a - $Z_{nS;}$ and Wall`s *B* and *Q* tests (Wall 1999) were calculated using DnaSP. Lower than expected values of Z_{nS} are caused by intragenic linked sites that have been subjected to a form of selection. ZZ can be used for detecting intragenic recombination (larger positive values of ZZ are expected with increasing recombination). Wall`s *B* and *Q* tests are used to detect balancing selection or population substructure, they do so by detecting events that produce trees with relatively longer external branches (Soriano 2008). Confidence intervals (95%) for these test statistics were estimated by coalescent-based simulations using DnaSP (Librado and Rozas 2009).

2.5.2.4.4 Nucleotide diversity *(π* and *θ*)*

Theta and it`s estimates

Nucleotide diversity is used to measure the degree of polymorphism within a population. One measure of nucleotide diversity is Theta.

Theta (*θ)* is known as the population mutation parameter. It is estimated from the observed number of polymorphic sites using the formula: $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the neutral mutation rate per locus per generation.

There are several methods of measuring *θ* from sequencing data (Theta estimators). These methods are calculated using different parameters derived from the observed diversity and making different assumptions (Jobling *et al.* 2004), including:

The number of alleles, the number of segregating sites (S), the number of singletons (ƞ) (see Sections 2.5.2.4.9 and 2.5.2.4.10, page 56 and 57, respectively), the observed homozygosity and the mean number of pairwise differences (π) (Jobling *et al.* 2004), These measures can be represented as θ_S. They should have the same value under neutral evolution.

Two of the previous measures of nucleotide diversity were used: The observed homozygosity (*θ*) and the mean number of pairwise differences (π). These two measures are calculated from the number of effectively silent sites in a sequence. π is calculated by the formula of Nei (1987): π = n(Σxixjπij)/(n-1), where n is the number of sequences, xi and xj the frequencies of the ith and jth sequences respectively and πij the proportion of different nucleotides between them. Standard deviation of π is the square root of the variance of π (Nei 1987) (see above for how to calculate θ). The test statistics used in this Thesis compare various Theta estimates based on the number of segregating sites, the pairwise number of differences and the number of derived alleles (see below).

2.5.2.4.5 Tajima`s *D* statistic*

Tajima`s *D* statistic (Tajima 1989) is used to test the neutral theory of evolution (Kimura 1983). *D* measures departure from neutrality reflected in the difference between low-and intermediate-frequency alleles. It is based on the fact that estimates of the number of segregating sites and the average number of nucleotide differences should be correlated under the neutral model.

2.5.2.4.6 Haplotype diversity and Fu`s *F*s statistic*

Haplotype diversity (H) is a measure of the uniqueness haplotypes in a given population. It is calculated as: H = (N/N-1) Σxi2, where N is the size of the sample and xi is the frequency of each haplotype in the sample (Nei and Tajima 1981). Fu`s *F*s statistic (Fu 1997) compares the observed number of haplotypes with the number expected under the assumption of an infinite-sites model of neutral mutation with no recombination.

2.5.2.4.7 Pairwise number of differences *# and raggedness (r) *

The pairwise number of differences shows the observed pairwise nucleotide differences (also called Roger and Harpending`s mismatch distribution), and the expected values in a population with constant size and no recombination and also with population growth-decline with free recombination (Rogers 1992).

The raggedness statistic is used to detect population expansion by quantifying the smoothness of the observed pairwise differences distribution (Harpending

1994). Under neutrality, the pairwise differences distribution has ragged peaks, which will be smoothed if the population has increased in size.

2.5.2.4.8 Allele-frequency spectrum*

The test compares the observed distribution of allelic frequency in a site (frequency spectrum) with the expected values in a constant size population (Tajima 1989).

2.5.2.4.9 Fu and Li`s *D** and *F** tests*

Fu and Li`s *D** and *F** test statistics (Fu and Li 1993) compare the observed number of singleton polymorphisms (η) with the number expected under a neutral model (Kimura 1983).

2.5.2.4.10 Fu and Li`s *D* and *F* test statistics with an outgroup*

Fu and Li`s *D* and *F* test statistics (Fu and Li 1993) are used to test the hypothesis that all mutations are selectively neutral (Kimura 1983). The *D* test statistic is based on the differences between ηe, the total number of mutations in external branches of the genealogy, and η, the total number of mutations (Fu and Li 1993), whereas the *F* test statistic is based on the differences between ηe, the total number of mutations in external branches of the genealogy, and k, the average number of nucleotide differences between pairs of sequences. Assuming the infinite sites model, DnaSP calculates the total number of mutations in the external branches of the genealogy as follows: at a given polymorphic site, the number of mutations in external branches is counted as the number of distinct singleton nucleotide variants (in the intraspecific data file) that are not shared with the outgroup. The total number of mutations in external branches of the genealogy is then calculated as the sum of the number of mutations in external branches of every polymorphic site. These tests require polymorphic data as well as data from an outgroup. Chimpanzee data was used as an outgroup.

2.5.2.4.11 Fay and Wu`s *H* test*

The *H* statistic (Fay and Wu 2000) measures departure from neutrality reflected in the difference between low-frequency and intermediate-frequency alleles. It is based on the differences between two estimators of θ: $\theta\pi$ (or k), the average number of nucleotide differences between pairs of sequences, and θ_H (Fay and Wu 2000), an estimator based on the frequency of the derived variants. The normalized *H* statistic is the scaled version of the *H* statistic (Zeng *et al*. 2006).

2.5.2.4.12 McDonald-Kreitman test (MKT)*

Under the neutral theory of evolution, the ratio of nonsynonymous to synonymous mutations within species (P_N/P_S) is expected to be equal to the ratio of nonsynonymous to synonymous mutations between species (D_n/D_s) also called (K_a/K_s) (see below). The McDonald-Kreitman test is used for the detection of signatures of selection (McDonald and Kreitman 1991). It compares the variation at synonymous and nonsynonymous sites within a species with the divergence between species.

2.5.2.4.13 K_a/K_s ratio*

The K_a/K_s ratio (or ω, dN/dS), is the ratio of the rate of nonsynonymous substitutions (K_a) to the rate of synonymous substitutions (K_s). This ratio can be used as an indicator of selection acting on a protein-coding gene. The ratio was calculated using different methods implemented in three software programs, DnaSP, Datamonkey and the K_a/K_s calculator. DnaSP uses the distance-based methods of Nei and Gojobori (Nei and Gojobori 1986). This method is a simpler version of the method developed by Miyata and Yasunaga (1980) (Miyata and Yasunaga 1980).

The advantage of this method over others is that it gives accurate results even when the per site number of nucleotide substitution is small. Datamonkey implements counting methods that count the number of nonsynonymous and synonymous substitutions: the fixed effect-likelihood (FEL) method estimates the K_a/K_s ratio on a site-by-site basis and the random-effect-likelihood method (REL) assumes K_a/K_s rates distributes across sites and then infers the rate at which sites evolve (Kosakovsky *et al*. 2005), whereas the K_a/K_s calculator (Zhang e*t al.* 2006b) uses various methods, of these, the NG algorithm (Jukes-Cantor Models) and maximum-likelihood methods were used.

The NG algorithm by Jukes and Cantor (1969) (Jukes and Cantor 1969) uses the following formula for calculating K_a/K_s: $d = 3/4 \log_e (1\text{-}4/3p)$, where p is either p_S or p_N. The maximum-likelihood method calculates nonsynonymous (K_a) and synonymous (K_s) substitution rates by model selection and model averaging. This method adopts the Akaike information criterion to measure fitness of data and models. This method aims to accurately capture evolutionary patterns in DNA-coding sequences (Zhang *et al.* 2006b).

2.5.2.5 Testing for purifying selection

2.5.2.5.1 Gene diversity

Gene diversity (H) at its most elementary level is represented by differences in nucleotide sequences. In diploid organisms it is equivalent to heterozygosity. It was estimated according to the formulae of Nei (1987): $H = n(1 - \Sigma x_i^2)/(n - 1)$, where n is the number of gene copies and x_i is the frequency of the ith allele (Nei 1987).

A web-based server called SMOGD, v1.2.5 was used to estimate genetic diversity (http://www.ngcrawford.com/django/jost/). SMOGD (Software for the measurement of genetic diversity) calculates the measures of genetic differentiation described by Hedrick (2005) and Jost (2008) (Hedrick 2005; Jost 2008). Bootstrap calculations of 95% confidence intervals were used to test the significance.

2.5.2.5.2 Genetic distance (*F_{ST}*)#

Wright`s fixation index (*F_{ST}*) is a measure of population differentiation, genetic distance based on differences in SNP allele frequencies in samples. *F_{ST}* is the correlation of randomly chosen alleles within a sub-population relative to the entire population. A common definition given is by Hudson *et al*. (1992) (Hudson *et al.* 1992): F_{ST} = ΠBetween - ΠWithin/ ΠBetween, where ΠBetween and ΠWithin represent the average number of pairwise differences between two individuals sampled from the same or different populations.

2.5.2.6 Time to the most recent common ancestor ($\mathrm{T_{MRCA}}$)

2.5.2.6.1 GeneTree

The GeneTree analysis was performed by Dr Charles Allerston of the University of Oxford, UK.

GeneTree, v.9.0 (http://www.stats.ox.ac.uk/griff/software.html) was used to estimate the time to the most recent common ancestral sequence ($\mathrm{T_{MRCA}}$) and the ages of some of the mutations. GeneTree uses maximum-likelihood coalescent analysis (Griffiths and Tavare 1997). The analysis is based on a coalescent model, and assumes an infinite-sites model of mutation with no recombination. Under this model, it is assumed that mutations occur along ancestral lineages, following a Poisson rate of $\theta/2$. θ is the population mutation rate. θ is calculated using the following formula: $\theta = 4\ N_e\mu$, where N_e is the effective population size and μ is the neutral mutation rate per region per generation. The neutral mutation rate μ can be calculated using: $\mu = \upsilon gL$, where υ is the neutral mutation rate per nucleotide per year, g is the generation time in years and L is the number of silent sites in the sequence.

3. Results and Discussion

Because of the implications of expressing full-length functional FMO2 (e.g, altered therapeutic drug metabolism), it is important to investigate whether any of the full-length alleles (23238C) are functional (i.e, do not have any nonsynonymous SNPs associated with them), and if so, at what frequency?

This was investigated by typing African individuals for six SNPs that were at high frequency in Africa, including the SNP g.23238C>T (Q472X), which causes the truncation of human FMO2. These SNPs were chosen on the basis of their high frequency in Africa and it is important to determine whether any of these SNPs occur on a full-length background, because of the possible effect on function.

3.1 Typing African individuals for six high-frequency SNPs

A total of 689 African samples from nine population groups (Table 4) (Bulsa n=90, Mambila n=48, Fulbe n=52, Manjak n=94, Sena n=83, Anuak n=109, Nuer n=80, Gurage n=83 and Chagga n=50) were genotyped for six SNPs that are at high frequency in African populations (Table 5), by TaqMan assay (see Section 2.3.4).

Table 4. The nine African population groups

Population Group	Name of Country	Number of Individuals
Fulbe	Cameroon	53
Mambila		50
Anuak	Ethiopia	109
Nuer		80
Gurage		83
Sena	Mozambique	83
Manjak	Senegal	94
Chagga	Tanzania	50
Bulsa**	Ghana	90

Table shows the nine population groups from Africa that were genotyped for six high-frequency SNPs. It also gives the number of individuals in each group as well as the country of origin. Population groups from the same country are color marked.

Table 5. The six SNPs and their frequency in Africa

dbSNP rs # cluster id	Position and amino acid change	Frequency in the HapMap-Yoruba panel
rs2020870	g.107A>G (D36G)	0.10
rs2020860	g.7731T>C (F81S)	0.28
rs28369860	g.10951delG (V113fsX)	0.21
rs2307492	g.13693T>C (F182S)	0.17
rs2020862	g.13732C>T (S195L)	0.56
rs6661174	g.23238C>T (Q472X)	0.83

Table displays the dbSNP rs # cluster id, the position in the gene, the amino acid change and the SNP frequency in the HapMap-Yoruba panel.

The genotyping results are shown in Table 6.

3.1.1 Genotypes

3.1.1.1 Hardy-Weinberg equilibrium (HWE)

Genotypes were tested for deviation from the HWE (Table 6). Genotyping revealed that three of the SNPs g.7731T>C (F81S), g.13693T>C (F182S) and g.13732C>T (S195L) are in HWE in all nine of the population groups, whereas the other three SNPs are in HWE only in some of the population groups. The significance of the deviation of genotype frequency from HWE was examined using Pearson`s chi-square test. There is one degree of freedom for the "within-a-group" test and eight degrees of freedom for the between-groups" test. The 5% significance level is 3.84 for one degree of freedom and 15.51 for eight degrees of freedom. Significant deviation from HWE is observed for the SNP g.107A>G (D36G) in the Mambila (*P*<0.001), Fulbe (*P*<0.001) and the Gurage (*P*<0.005).

The deviation from HWE in Mambila is caused by a higher number of observed heterozygote genotypes, whereas in the Fulbe and Gurage it is caused by fewer than expected heterozygote genotypes. The SNP g.10951delG (V113fsX) deviated significantly from HWE in Manjak (*P*<0.005), Gurage (*P*<0.001) and in all population groups combined (*P*<0.005) due to fewer than expected observed heterozygotes.

The SNP g.23238C>T (Q472X) deviated significantly from HWE due to more than expected heterozygotes in Manjak (*P*<0.05), Nuer (*P*<0.005) and in the Gurage, although not significantly (*P*>0.05).

Most of the deviation from HWE was because of fewer than expected observed heterozygote genotypes. The presence of fewer heterozygote genotypes in these population groups may be due to inbreeding, which leads to a reduction in heterozygotes and an increase in homozygotes (Charlesworth and Charlesworth

1987). In an attempt to test whether these results are confined to any geographic location, west and east-African genotyping results were pooled and Pearson`s chi square test was used to test for any significant differences in the number of derived SNP alleles between west and east Africa. Chi square, (χ_2 = 1.8) was not significantly different between west and east-African populations (P<0.2) (results not shown).

Table 6. Genotyping results for the nine population groups

Ethnic	**Genotyping**	**g.107A>G D36G**	**g.7731T>C F81S**	**g.10951delG V113fsX**	**g.13693T>C F182S**	**g.13732C>T S195L**	**g.23238C>T Q472X**
Group		**A/A A/G G/G**	**T/T T/C C/C**	**G/G G/- -/-**	**T/T T/C C/C**	**C/C C/T T/T**	**C/C C/T T/T**
Bulsa	**Observed**	**66 9 0 (*n*=75)**	**46 39 4 (*n*=89)**	**44 29 11 (*n*=84)**	**57 29 3 (*n*=89)**	**23 41 23 (*n*=87)**	**3 21 66 (*n*=90)**
(n=90)	**Chi-square**	**0.13**	**1.45**	**2.83**	**0.09**	**0.29**	**0.65**
Mambila	**Observed**	**9 29*** 1 (*n*=39)**	**28 14 5 (*n*=47)**	**29 14 5 (*n*=48)**	**36 11 1 (*n*=48)**	**12 22 12 (*n*=46)**	**7 20 23 (*n*=50)**
(n=48)	**Chi-square**	**11.91*** **	**2.24**	**2.37**	**0.02**	**0.09**	**0.59**
Fulbe	**Observed**	**27 6***7 (*n*=40)**	**30 19 3 (*n*=52)**	**30 18 4 (*n*=52)**	**36 16 0 (*n*=52)**	**14 19 15 (*n*=48)**	**0 18 34 (*n*=52)**
(n=52)	**Chi-square**	**14.4*** **	**0**	**0.31**	**1.72**	**2.08**	**2.28**
Manjak	**Observed**	**71 5 1 (*n*=77)**	**40 40 10 (*n*=90)**	**36 30** 23 (*n*=89)**	**55 35 3 (*n*=93)**	**18 35 19 (*n*=72)**	**6 22* 66 (*n*=94)**
(n=94)	**Chi-square**	**3.05**	**0**	**8.62** **	**0.84**	**0.05**	**4.15* **
Nuer	**Observed**	**49 17 0 (*n*=66)**	**37 36 7 (*n*=80)**	**62 18 0 (*n*=80)**	**50 26 3 (*n*=79)**	**25 38 17 (*n*=80)**	**11 20** 48 (*n*=79)**
(n=80)	**Chi-square**	**1.44**	**0.18**	**1.29**	**0.03**	**0.13**	**9.75** **

Gurage	**Observed**	**53 13** 6 (*n*=72)**	**38 34 11 (*n*=85)**	**66 12*** 5 (*n*=83)**	**75 6 1 (*n*=82)**	**22 40 20 (*n*=82)**	**14 30 38 (*n*=82)**
(n=83)	**Chi-square**	**9.9****	**0.58**	**11.44*****	**3.67**	**0.05**	**3.27**
Anuak	**Observed**	**66 27 4 (*n*=97)**	**55 43 9 (*n*=107)**	**84 24 1 (*n*=109)**	**66 34 8 (*n*=108)**	**32 57 19 (*n*=108)**	**5 8 55 (*n*=108)**
(n=109)	**Chi-square**	**0.33**	**0.02**	**0.25**	**1.43**	**0.55**	**1.86**
Chagga	**Observed**	**38 4 0 (*n*=42)**	**19 23 6 (*n*=48)**	**34 15 1 (*n*=50)**	**31 17 2 (*n*=50)**	**14 25 10 (*n*=49)**	**2 18 30 (*n*=50)**
(n=50)	**Chi-square**	**0.11**	**0.06**	**0.2**	**0.03**	**0.04**	**0.12**
Sena	**Observed**	**57 8 0 (*n*=65)**	**43 33 5 (*n*=81)**	**47 29 6 (*n*=82)**	**43 37 3 (*n*=83)**	**20 43 13 (*n*=76)**	**3 23 56 (*n*=82)**
(n=83)	**Chi-square**	**0.28**	**0.16**	**0.27**	**2.16**	**1.52**	**0.11**
Combined	**Observed**	**435 118 14 (*n*=567)**	**336 281 60 (*n*=677)**	**432 189* 56 (*n*=677)**	**449 211 24 (n=684)**	**180 320 148 (*n*=648)**	**51 220 416 (*n*=687)**
(n=691)	**Chi-square**	**2.97**	**0.01**	**25.11***	**0.02**	**0.06**	**7.96**

n in the ethnic group column represents the total number of samples from that ethnic group, whereas the n (italics) in the other columns represents the number of individuals of each group who were successfully genotyped for a particular SNP.
Cases in which a SNP genotype departs significantly from HWE are indicated in red. **P*<0.05, ***P*<0.005 and ****P*<0.001.

3.1.2 SNP frequencies

Frequencies of the SNPs are shown in Table 7. All the frequency values are for the derived allele. The g.23238C>T (Q472) SNP occurred at high frequency in the combined samples (0.77), followed by g.13732C>T (S195L), with a relatively high frequency (0.48), whereas the other four SNPs occurred at a relatively low frequency, 0.14 and 0.19 respectively for g.107A>G (D36G) and g.13693T>C (F182S), the g.10951delG (V113fsX) and the g.7731T>C (F81S) SNPs had an intermediate frequency of 0.23 and 0.30, respectively.

The frequencies of three of the SNPs, g.107A>G (D36G), g.10951delG (V113fsX) and g.13693T>C (F182S)), are significantly different (*P<0.001*), (*P<0.001*) and (*P<0.05*), respectively, among populations. The g.107A>G (D36G) SNP occurs at a higher frequency in the three Ethiopian groups: Nuer, Gurage and Anuak, at 0.23, 0.17 and 0.17, respectively, and at an even higher frequency (0.40) in a group from Cameroon, the Mambila, whereas the g.10951delG (V113fsX) SNP has a higher frequency in five of the population groups, Bulsa (0.30), Mambila (0.25), Fulbe (0.32), Manjak (0.43) and Sena (0.25). For the g.13693T>C (F182S) SNP, the frequency is significantly lower in the Gurage (0.05).

The frequency of the other three SNPs [g.7731T>C (F81S), (F182S), g.13732C>T (S195L) and g.23238C>T (Q472X)] was not significantly different.

The frequencies of all six SNPs in the combined African populations are similar to the frequency reported for the NCBI database in the African-HapMap Panel. The frequency of the g.107A>G (D36G) SNP in the combined African populations is 0.14, compared with 0.10 in the panel; the frequency of the g.7731T>C (F81S) SNP in the combined African population is 0.30, compared with 0.28 in the panel, the frequency of the g.10951delG (V113fsX) SNP in the combined African populations is 0.23, compared with 0.21 in the panel; the frequency of the g.13693T>C (F182S) SNP in the combined African population is 0.19, compared with 0.17 in the panel, the g.13732C>T (S195L) SNP frequency in the combined African populations is 0.48, compared with 0.56 in the panel; and the frequency of the g.23238C>T (Q472X) SNP is 0.77 in the combined African populations compared with 0.83 in the panel.

The results show that these SNPs are not distributed evenly across Africa, which might have important implications for the frequency of potentially functional FMO2 across sub-Saharan Africa, depending on whether SNPs occur on a full-length, 23238C allele or on a truncated, non-functional, 23238T allele.

3.1.3 Haplotype inference

Haplotypes were inferred for each of the nine African population groups, but only for individuals who had been successfully typed for all of the six SNPs. Phase, version 2.1 (see Section 2.5.1.1), both with and without recombination was used for inferring haplotypes. Only inferred haplotypes with a confidence value of >95% were chosen. The number of high-confidence inferred haplotypes obtained was greater when using Phase without recombination than with recombination. Therefore results from the former were used.

Table 8 shows the haplotypes inferred from the Phase analysis of the genotypes, and identifies the nucleotide sequence as well as the amino acid residues the nucleotides code for.

The ancestral haplotype at each of the six SNP positions was inferred by comparison of *FMO2* SNP sequences with the corresponding positions in *FMO2* of chimp and mouse. At all six positions, one of the SNP variants was identical to the nucleotide present in the *FMO2* of both chimpanzee and mouse and, thus, was considered to be the ancestral form. The inferred ancestral sequence at the six SNP positions referred to in Table 8, the pie charts and Network diagrams is 5'ATGTCC3'.

Table 7. SNP frequencies in the African populations

	SNP Frequencies											
	rs2020870		rs2020860		rs28369860		rs2307492		rs2020862		rs6661174	
	g.107A>G (*D36G)*		*g.7731T>C* (*F81S)*		*g.10951delG* (V113fsX)		*g.13693T>C* (F182S)		*g.13732C>T* (S195L)		*g.23238C>T* (Q472X)	
Ethnic Group	*No of individuals Genotyped (n)*	*No & frequency of derived allele*	*No of individuals genotyped (n)*	*No & frequency of derived allele*	*No of individuals genotyped (n)*	*No & frequency of derived allele*	*No of individuals genotyped (n)*	*No & frequency of derived allele*	*No of individuals genotyped (n)*	*No & frequency of derived allele*	*No of individuals genotyped (n)*	*No & frequency of derived allele*
Bulsa	74	9 (0.06)	89	47 (0.26)	84	51 (0.30)	89	35 (0.20)	87	87 (0.50)	90	153 (0.85)
Mambila	39	31** (0.40)	47	24 (0.26)	48	24 (0.25)	48	13 (0.14)	46	46 (0.50)	50	66 (0.66)
Fulbe	40	10 (0.14)	52	25 (0.24)	52	36 (0.32)	52	16 (0.15)	48	49 (0.51)	52	86 (0.83)
Manjak	77	7 (0.05)	90	60 (0.33)	89	76* * (0.43)	93	41 (0.22)	72	73 (0.51)	94	154 (0.82)
Nuer	66	35 (0.23)	80	50 (0.31)	80	18 (0.11)	79	32 (0.22)	80	72 (0.45)	79	116 (0.73)
Gurage	72	25 (0.17)	85	56 (0.34)	83	22 (0.13)	82	8* (0.05)	82	80 (0.49)	82	106 (0.64)
Anuak	97	35 (0.18)	107	61 (0.29)	109	26 (0.12)	108	50 (0.23)	108	95 (0.44)	108	158 (0.73)
Chagga	42	4 (0.05)	48	35 (0.36)	50	17 (0.17)	50	21 (0.21)	49	45 (0.46)	50	78 (0.78)
Sena	65	8 (0.06)	81	43 (0.27)	82	41 (0.25)	83	43 (0.26)	76	69 (0.45)	82	135 (0.82)
Combined	572	164 (0.14)	679	401 (0.30)	677	311 (0.23)	684	259 (0.19)	648	616 (0.48)	687	1052 (0.77)
Chi^2 (df=8)	*37.3***		*1.8*		*31.3***		*16.5**		*3.64*		*1.43*	

n=number of individuals genotyped. 2n (number of alleles genotyped) was used in calculating the frequency of the derived allele for each SNP. SNP, single-nucleotide polymorphism. Significant differences in the derived allele are marked with * (*P*<0.05) and **(*P*<0.001).

Table 8. Haplotypes inferred from Phase analysis of genotypes

Haplotype Number	DNA	Protein
1 (Ancestral)	ATGTCC	DFVFSQ
2	ATGCCT	DFVSSstop
3	AC-TTT	DS-FLstop
4	ATGTTT	DFVFLstop
5	ATGTCT	DFVFSstop
6	GTGTTT	GFVFLstop
7	AT-TTT	DF-FLstop
8	ACGTCT	DSVFSstop
9	ACGTTT	DSVFLstop
10	ACGTCC	DSVFSQ
11	GCGTTT	GSVFLstop
12	ATGCCC	DFVSSQ
13	ATGTTC	DFVFLQ
14	ACGCCT	DSVSSstop
15	AT-CCT	DF-SSstop
16	AT-TCC	DF-FSQ
17	GTGTTC	GFVFLQ
18	GCGTTC	GSVFLQ
19	AC-TCT	DS-FSstop
20	GCGTCC	GSVFSQ
21	GC-TTT	GS-FLstop
22	AT-TTC	DF-FLQ
23	GCGTTC	GSVFLQ
24	GTGTCC	GFVFSQ
25	GTGCCT	GFVSSstop
26	AC-TTC	DS-FLQ
27	ACGCCC	DSVSSQ
28	GT-TCC	GF-FSQ
29	GTGTCT	GFVFSstop
30	AT-TCT	DFVFSstop

Haplotype frequencies were estimated by counting.

Table 9 shows the frequency of the Phase-inferred haplotypes in the nine African populations. The haplotype frequencies were tested for significant differences among population groups. The significantly different values are marked with an asterix. It was not possible to calculate the chi-square values for some of the haplotypes, since there were many zero values and the expected values were below 5. For others, such as haplotype 2, groups with zero were excluded and the degrees of freedom adjusted accordingly. In Table 9, the haplotypes are listed from top to bottom, according to their frequency in the total population sample.

The results showed that six haplotypes, 1 (ancestral), 2, 3, 4, 5 and 6, are present at relatively high frequency in some of the population groups. Haplotypes 1 (ancestral), 4 and 5 are present in all groups, but at different frequencies. Haplotype 2 is not observed in Gurage, whereas haplotype 3 is not observed in Nuer and haplotype 6 is not observed in Fulbe.

Another five haplotypes are present at moderate frequency in some population groups: haplotypes 7, 8, 9, 10 and 11. These eleven haplotypes make up 0.93 of all haplotypes observed. A few haplotypes were observed in only one of the groups: haplotype 17 in Gurage, haplotype 18 in Nuer and haplotype 20 in Anuak, whereas other haplotypes were individual-specific: haplotypes 22 and 23 in Mambila, haplotype 24 in Sena, haplotypes 25 and 29 in Fulbe, haplotype 26 in Sena, haplotype 27 in Nuer, haplotype 28 in Manjak and haplotype 30 in Bulsa.
All of the group–specific haplotypes occur at a low frequency.

Pie charts displaying the haplotype frequencies in each of the population groups are shown in Figure 12. The charts indicate that there are differences in haplotype frequencies among the groups. The haplotypes have been sorted according to frequency.

The frequencies of haplotypes 1 (ancestral), 2, 3, 5 and 6 were significantly different between groups. The significant difference in haplotype 1 (ancestral) is due to the higher than expected frequency in Mambila (0.35, $P<0.05$), for haplotype 2 it is caused by the absence of the haplotype in Gurage and by lower than expected frequency in Mambila (0.06, $P<0.01$), whereas for haplotype 3 it is due to the frequency ranging from as low as 0.00 in Nuer and 1 and 0.02 in Anuak and Chagga respectively, to a higher than expected frequency of 0.29 in Manjak ($P<0.001$). The significant difference in the frequency of haplotype 5, however, is due to the higher than expected frequency in Gurage and Nuer, 0.24 and 0.21, respectively ($P<0.001$). Haplotype 6 has a significantly higher than expected frequency in Anuak (0.15, $P<0.001$). The results indicate that haplotype 1 (ancestral) has a relatively high frequency in all population groups, which suggests that the frequency of potentially functional FMO2 may be high across sub-Saharan Africa. Interestingly, the other haplotypes that had a significantly different frequency, haplotypes 2, 3, 5 and 6 are all non-functional. These differences in non-functional haplotype frequencies may be explained by the 23238T allele accumulating other mutations in different regions and therefore resulting in these differences.

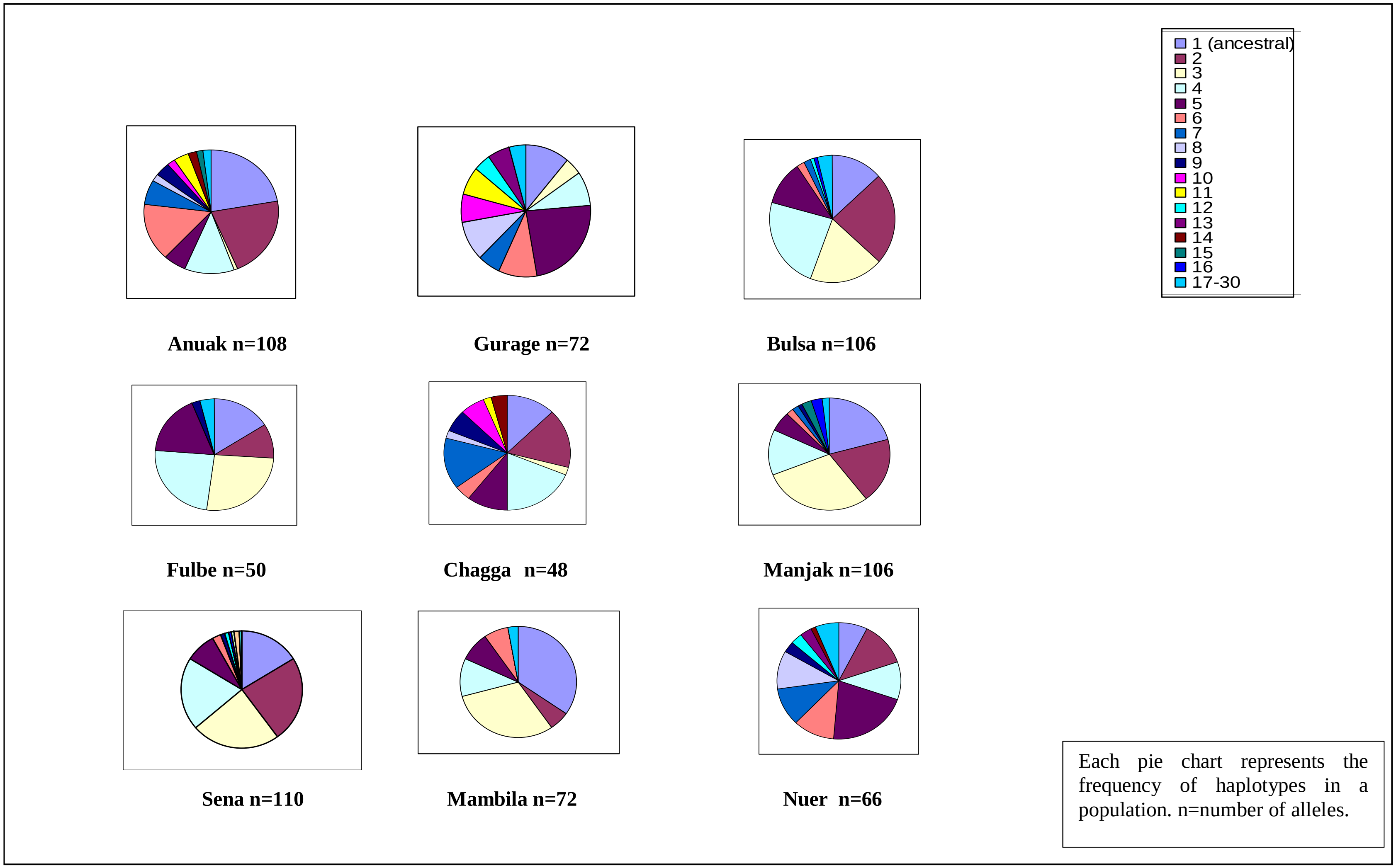
Table 9. Haplotype number and frequency in the African Populations

Haplotype ID	**Anuak (n=108)**		**Gurage (n=72)**		**Bulsa (n=106)**		**Fulbe (n=50)**		**Chagga (n=48)**		**Manjak (n=106)**		**Sena (n=110)**		**Mambila (n=72)**		**Nuer (n=66)**		***Total (n=738)***		***Chi Square***
1	24	0.22	8	0.11	14	0.13	8	0.16	6	0.13	22	0.21	18	0.16	25*	0.25	5	0.08	n=130	0.176	16.4*
2	23	0.22	0*	0.00	25	0.24	5	0.10	8	0.17	20	0.19	26	0.24	4*	0.06	8	0.12	n=119	0.161	24**
3	1*	0.01	3	0.04	20	0.19	13	0.26	1*	0.02	31*	0.29	26	0.24	22	0.31	0*	0.00	n=117	0.158	54.9***
4	13	0.12	6	0.08	25	0.24	12	0.24	9	0.19	14	0.13	22	0.20	8	0.11	7	0.11	n=116	0.157	7
5	6	0.06	17*	0.24	12	0.11	9	0.18	5	0.10	6	0.06	9	0.08	6	0.08	14*	0.21	n=84	0.114	42***
6	16 *	0.15	7	0.10	2	0.02	0	0.00	2	0.04	2	0.02	3	0.03	5	0.074	7	0.11	n=44	0.06	39.6***
7	7	0.06	4	0.06	2	0.02	0	0.00	7	0.15	2	0.02	0	0.00	0	0.00	7	0.00	n=29	0.04	N/A
8	2	0.02	7	0.10	0	0.00	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	7	0.11	n=17	0.23	N/A
9	4	0.04	0	0.00	0	0.00	1	0.02	3	0.06	1	0.01	1	0.01	0	0.00	2	0.03	n=12	0.16	N/A
10	2	0.02	5	0.07	0	0.00	0	0.00	3	0.06	0	0.00	0	0.00	0	0.00	0	0.00	n=10	0.13	N/A
11	4	0.04	5	0.07	0	0.00	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	n=10	0.13	N/A
12	0	0.00	3	0.04	1	0.01	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	2	0.03	n=7	0.01	N/A
13	0	0.00	4	0.06	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.03	n=6	0.008	N/A
14	2	0.02	0	0.00	0	0.00	0	0.00	2	0.04	0	0.00	0	0.00	0	0.00	1	0.02	n=5	0.007	N/A
15	2	0.02	0	0.00	0	0.00	0	0.00	0	0.00	3	0.03	0	0.00	0	0.00	0	0.00	n=5	0.007	N/A

16	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	3 0.03	1 0.01	0 0.00	0 0.00	n=5 0.007	N/A
17	0 0.00	3 0.04	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	n=3 0.004	N/A
18	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	3 5%	n=3 0.004	N/A
19	0 0.00	0 0.00	2 0.02	0 0.00	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	n=3 0.004	N/A
20	2 0.02	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	n=2 0.003	N/A
21	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	0 0.00	n=2 0.003	N/A
22	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.01	0 0.00	n=1 0.001	N/A
23	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.01	0 0.00	n=1 0.001	N/A
24	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	n=1 0.001	N/A
25	0 0.00	0 0.00	0 0.00	1 0.02	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	n=1 0.001	N/A
26	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	n=1 0.001	N/A
27	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.02	n=1 0.001	N/A
28	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	0 0.00	n=1 0.001	N/A
29	0 0.00	0 0.00	0 0.00	1 0.02	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	n=1 0.001	N/A
30	0 0.00	0 0.00	1 0.01	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	n=1 0.001	N/A

Haplotype 1 is the ancestral haplotype, n=number of alleles. Haplotypes are listed from top to bottom according to frequency. Significant differences are marked with * *P*<0.05, ***P*<0.01, ****P*<0.001.

Figure 12. Phase-inferred Haplotype Frequency Pie Charts



Each pie chart represents the frequency of haplotypes in a population. n=number of alleles.

3.1.4 Impact of SNPs on a23238C allele background

Because of the implications of expressing functional FMO2 (i.e., altered drug metabolism), it is important to estimate the proportion of 23238C alleles that are associated with a nonsynonymous SNP. Studies have shown that two of the five SNPs, g.10951delG (V113fsX) and g.13732C>T (S195L), cause loss of function (Furnes *et al*. 2003; Krueger *et al*. 2005), whereas the effect of the other three is unknown.

Table 10 shows the number of full-length, 23238C, and truncated, 23238T, alleles. For each of the five SNPs, the proportion of truncated alleles on which the SNP occurs is greater than the proportion of full-length alleles.

It is expected that the truncated 23238T alleles will accumulate more nonsynonymous mutations because it is a null-allele and therefore there is a relaxation in the selective constraint (no purifying selection).

The results indicate that the frequency of potentially functional FMO2 may be high across sub-Saharan Africa. The frequencies displayed in Table 11 are the percentage of full-length, 23238C, and truncated, 23238T, alleles on which a derived SNP allele is present. For this calculation, haplotypes of individuals who had been genotyped for all six SNPs were inferred by Phase (see Tables 8 and 9). To examine the percentage of full-length 23238C, and truncated, 23238T alleles on which one of the derived alleles of the five other SNPs [(g.107A>G (D36G), g.7731T>C (F81S), g.10951delG (V113fsX), g.13693T>C (F182S) and g.13732C>T (S195L)] occur, first the total number and proportion of the derived alleles for each SNP were calculated by adding all haplotypes that contain the derived SNP allele and dividing the resulting number by the total number of successfully typed alleles. Second the number of full-length 23238C alleles that have a SNP present was counted using the haplotype data in Table 9, and the proportion is calculated by dividing the number of full-length 23238C alleles on which the derived allele for a SNP is present by the number of total full-length 23238C alleles. This was repeated for each of the five SNPs. For the g.107A>G (D36G) SNP for example, the total number of derived alleles is 69, this number is the sum of all 11 haplotypes (6, 11, 17, 18, 20, 21, 23, 24, 25, 28 and 29) that contain the derived allele of the SNP. The proportion is calculated by dividing 69 by 738, which is the total number of successfully typed alleles. Out of the 69 derived alleles, 11 are present on full-length 23238C alleles. The proportion is calculated by dividing 11 by the total number of full-length alleles, 171. The same is repeated for the truncated, 23238T alleles.

Table 10. Number of full-length and truncated haplotypes in each population group

Population group										
Haplotype	Anuak	Gurage	Bulsa	Fulbe	Chagga	Manjak	Sena	Mambila	Nuer	All
23238C Full-length	28	23	16	8	9	26	22	26	13	171
23238T Truncated	80	49	90	42	39	80	88	46	53	567
Total	108	72	106	50	48	106	110	72	66	738

Numbers display the number of haplotypes in each population group that have a C or T at the g.23238C>T (Q472X) locus.

Table 11. Proportion of 23238C and 23238T alleles on which a nonsynonymous SNP is present

SNP Amino acid change	Effect on function	Total number and proportion of derived allele		Number and proportion of full-length alleles on which SNP is present		Number and proportion of truncated alleles on which SNP is present	
g.107A>G (D36G)	(Effect unknown)	69	0.09	11	0.06	58	0.10
g.7731T>C (F81S)	(Effect unknown)	184	0.25	18	0.11	166	0.29
g.10951delG (V113fsX)	(Loss of function)	164	0.22	8	0.05	156	0.28
g.13693T>C (F182S)	(Effect unknown)	138	0.19	8	0.05	130	0.23
g.13732C>T (S195L)	(Loss of function)	345	0.47	8	0.05	337	0.59

Total number of 23238C alleles=171
Total number of 23238T alleles=567
Total number of alleles=738.

The results show that each of the five SNPs can occur on a 23238T or a 23238C allele background. However, for each SNP, the proportion of 23238C haplotype backgrounds on which it occurs is relatively low (0.05-0.11).

In an attempt to estimate the frequency of potentially functional FMO2 in each of the nine population groups (i.e., haplotypes that contain none of the other five SNPs i.e., the ancestral haplotype) the frequency of total 23238C alleles (ancestral + all others with nonsynonymous SNPs associated) was estimated for each population group (Table 12). In Anuak, for example, the total number of alleles (both full-length 23238C and truncated 23238T alleles) was 108, the total number of full-length 23238C alleles (both functional ancestral and full-length with nonsynonymous SNPs associated) was 28 and out of these, 24 were on ancestral haplotypes, i.e, contained none of the five nonsynonymous SNPs.

To estimate the frequency of full-length 23238C alleles, the total number of 23238C alleles (28) was divided by the number of total alleles (108), which resulted in a frequency of 0.26. This was repeated for each of the nine population groups. As for the numbers and frequencies in the third column of Table 12, they represent the number of ancestral (functional) 23238C alleles and the frequency was calculated by dividing the number of functional 23238C alleles (24), by the number of total 23238C alleles (28), which resulted in a frequency of 0.86.

Table 12. Frequency of expected potentially functional FMO2 in the African Populations

Population Group	Number of total alleles	Number of full-length C alleles and their frequency	Number of C alleles that encode a potentially functional FMO2		
				Frequency of full-length	Frequency of total
Anuak	108	28 [0.18-0.34] (0.26)	24	(0.86)	(0.22)
Gurage	72	23 [0.21-0.43] (0.32)	8	(0.35)	(0.11)
Bulsa	106	16 [0.08-0.22] (0.15)	14	(0.88)	(0.13)
Fulbe	50	8* [0-0.06] (0.02)	8	(1.00)	(0.16)
Chagga	48	9 [0.08-0.30] (0.19)	6	(0.67)	(0.13)
Manjak	106	26 [0.17-0.33] (0.25)	22	(0.85)	(0.21)
Sena	110	22 [0.12-0.27] (0.20)	18	(0.82)	(0.16)
Mambila	72	27 [0.27-0.49] (0.38)	25	(0.93)	(0.35)
Nuer	66	13 [0.10-0.30] (0.20)	5#	(0.38)	(0.08)
In all groups combined	738	172 (0.23)	130	(0.76)	(0.18)

*Chi-square=25.4, *P*<0.001. DF=8
#Chi-square=35.7, *P*<.0.001. DF=8
DF, Degrees of freedom.
Numbers in brackets are the 95% confidence intervals calculated using an online calculator available at: http://www.dimensionresearch.com/resources/calculators/conf_prop.html.

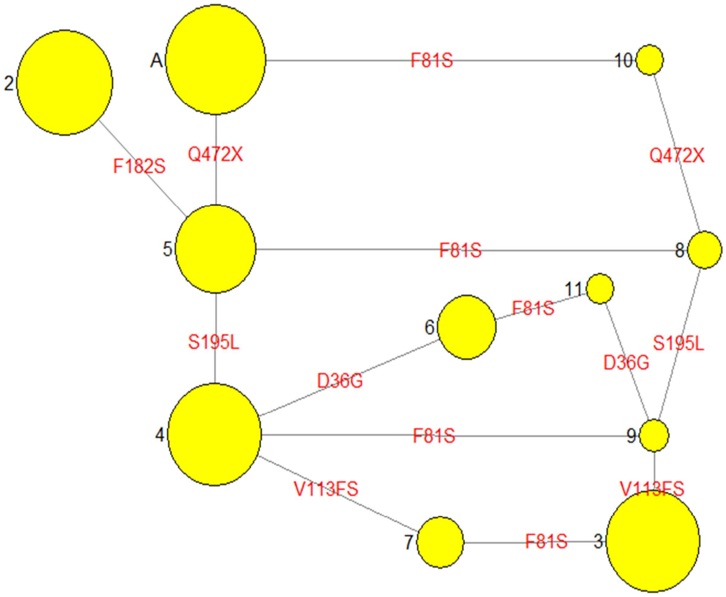
The groups with the highest frequency of potentially functional 23238C alleles, compared with the total number of alleles in a population, were Mambila, Anuak and Manjak, with a frequency of 0.35, 0.22 and 0.21, respectively. The group with the lowest frequency was Nuer, with a frequency of 0.08. The proportion of the full-length (23238C) alleles in each group that is expected to be potentially functional was observed to be high in most of the groups, reaching values of 1.00 in Fulbe and 0.93 in Mambila (Table 12). It was interesting to note that Gurage and Nuer had a relatively low frequency, 0.35 and 0.38, respectively.

Recombination could result in the five SNPs occurring on a 23238C allele background in these groups, which might explain the relatively low frequency of potentially functional FMO2 observed in some of these groups. The number of 23238C alleles that encode a potentially functional FMO2 is significantly different among groups (*P<0.001*). The frequency of potentially functional FMO2, in comparison with total alleles in each population group, was significantly different among groups (*P<0.001*) (Table 12). The results show that the frequency of potentially functional FMO2 is relatively similar in west and east Africa, 0.19 and 0.15, respectively (results not shown).

These results are similar to those of Veeramah *et al*. (Veeramah *et al*. 2008), in regards to the frequency of the *FMO2*1* allele being high and relatively uniform across sub-Saharan Africa, but, interestingly, in contrast, the frequency of potentially functional 23238C haplotypes is significantly different across sub-Saharan Africa. This is an important finding since the study of Veeramah *et al*. only looked at the frequency of full-length 23238C alleles across sub-Saharan Africa, but did not determine whether these alleles have any nonsynonymous SNPs associated with them, which if present, would render the full-length alleles nonfunctional. My results indicate that a high proportion of full-length 23238C alleles are potentially functional since they do not have any of the five nonsynonymous SNPs associated. This was further examined by resequencing, to test for the presence of other nonsynonymous SNPs.

3.1.5 Visualization of haplotypes using Network

The inferred haplotypes were input into Network to gain insights into the mutational relationship of the SNPs. The reduced-median algorithm was used, since all the SNPs were in the binary form. An input file was created for each of the nine population groups separately and one with all the groups combined. Figure 13 displays the mutational relationship among haplotypes in the combined African groups. As the diagram shows, mutations occur on both an ancestral allele background (23238C) as well as on the 23238T allele background. Some population groups like the Anuak and Gurage, for example, have a more complicated Network diagram compared with the one for some of the other groups, like the Mambila, for example. This could be due to a number of reasons, including recurrent mutations, recombination and genotyping errors. Genotyping errors can be reasonably excluded, since the individuals in question fall into the middle of a called cluster. Furthermore, genotyping was repeated several times, with the same result.



Haplotype id	Haplotypes	Frequency
A	ATGTCC	130
2	ATGCCT	119
3	AC-TTT	117
4	ATGTTT	116
5	ATGTCT	84
6	GTGTTT	44
7	AT-TTT	29
8	ACGTCT	17
9	ACGTTT	12
10	ACGTCC	10
11	GCGTTT	10

Figure 13. Network diagram for the combined African Populations. This diagram indicates that one of the five nonsynonymous SNPs, g.7731 T>C (F81S) occurs on a 23238C background, whereas the other four occur on a 23238T background, i.e, after the g.23238C>T (Q472X) SNP. The identity and number of haplotypes are shown in the legend.

The results indicate that the frequency of potentially functional FMO2 is high in sub-Saharan Africa and that linked SNPs mostly occur on a 23238T allele background. The low frequency of occurrence of SNPs on a 23238C allele background could be explained by recombination causing mutations to be associated with a 23238C allele. The Network diagram (see Figure 13) indicates that the five nonsynonymous SNPs mostly occur on a 23238T allele background, i.e, after the g.23238C>T (Q472X) SNP, and that when these SNPs are observed on a 23238C allele background, it is at low frequency as observed in Figure 13, where the SNP g.107A>G (D36G) occurs on a 23238C allele background only once (haplotype 18). This is further supported by the SNPs occurring at high frequency on a 23238T allele background and if they do occur on a 23238C allele background, they do so at low frequency as a result of recombination.

It is important to determine whether any other nonsynonymous SNPs occur on a 23238C allele background, therefore resequencing was performed. In addition, this will provide insights into the evolutionary history of *FMO2* and will make available data for testing for evidence of natural selection at the *FMO2* locus.

3.2 *FMO* resequencing data

The genotype of individuals with respect to the g.23238C>T (Q472X) mutation had to be determined before performing resequencing since a sample of individuals had to be chosen that represents the proportion of 23238C and 23238T alleles in Africa. Individuals were chosen from two African (west and east) ascertainment plates (see Sections 2.2.2.1 and 2.2.2.2).

3.2.1 23238C→T genotyping

Of the 747 individuals in the two ascertainment panels, 336 had been previously genotyped for g.23238C>T (dbSNP #rs6661174) by Dr. Krishna Veeramah of UCL. The remaining 411 samples were genotyped by me using TaqMan assays (see Section 2.3.4.1 for methods and Table 13 for genotyping results). Genotyping results showed that for both ascertainment plates all three genotypes (23238C homozygotes, 23238T homozygotes and 23238C/T heterozygotes) are in HWE.

None of the three genotypes had a significantly different frequency when the two ascertainment plates were compared (P<0.4).

Of the 747 DNA samples from both west and east Africa, 12 homozygotes for the full-length *FMO2*1* allele (CCs) and 12 homozygotes for the truncated *FMO2*2A* allele (TTs) were chosen randomly from both the west- and east-African ascertainment plates. The resequencing samples do not reflect the frequency of the full-length (*FMO2*1*) allele in either west or east Africa. To reflect the overall frequency of the 23238C alleles in the west- and east- African panels (0.21 and 0.18, respectively, with an average of ~0.20), the number of 23238C alleles in the resequencing sample should be reduced to 6 alleles, in comparison with the 24 23238T alleles.

Table 13. 23238C→T genotyping results for individuals from African ascertainment plates

Ascertainment plate	Total number of individuals	23238C homozygotes	23238T homozygotes	23238 C/T heterozygotes	Undetermined
West Africa	**370**	**14**	**215**	**126**	**15**
East Africa	**377**	**10**	**243**	**114**	**10**

There is no significant difference in genotype frequencies between plates (*P*<0.4).

3.2.2 *FMO* nucleotide sequence variation

Forty-eight African DNA samples, 24 23238C and 24 23238T homozygotes, were resequenced for all eight coding exons and flanking intronic sequences of *FMO2.* In total, 4705 bp of the human *FMO2* gene were sequenced comprising 1605 bp of coding sequence, from exons 2 to 9 (535 codons, including the stop-codons in exons 6 and 9) and 3100 bp of intronic and 5`- and 3`-untranslated region.

The sequencing resulted in the identification of 32 variants. No novel SNPs were discovered. There were 8 transversions, 23 transitions and 1 insertion/deletion variants. Eighteen were in the coding region, of which 12 were nonsynonymous: in exon 2, g.107A>G (Asp→Gly change at amino acid residue 36; in exon 3, g.7695T>A (Phe→Tyr change at residue 69), g.7731T>C (Phe→Ser change at residue 81); in exon 5, g.13693T>C (Phe→Ser change at residue 182), g.13732C>T (Ser→Leu change at residue 195); in exon 6, g.18237G>A (Arg→Gln change at residue 238), g.18269C>T (Arg→Stop-codon at residue 249); in exon 7, g.19679A>G (Glu→Gly change at residue 314; g.19910G>C (Arg→Thr change at residue 391); in exon 8, g.22060T>G

(Asn→Lys change at residue 413); in exon 9, g.23115A>C (Asn→Thr change at residue 431), g.23238C>T (Gln→Stop-codon at residue 472).

One causes a frame shift: in exon 4, g.10951delG (Val→Stop-codon at residue 113), and 5 were synonymous: in exon 5, g.13733G>A (Ser→Ser at residue 195), in exon 7, g.19910G>C (Ala→Ala at residue 367), in exon 8, g.22027G>A (Glu→Glu at residue 402), in exon 9, g.230087A>G (Gly→Gly at residue 421) and g.23300A>G (Lys→Lys at residue 492).

For the combined samples, there were two singletons: g.18269C>T (R249X) in exon 6, and g.19679A>G (E314G) in exon 7, and three doubletons: g.7695T>C (F69Y) in exon 3, g.10951delG (V113fsX) in exon 4 and +18559T>C in intron 6. The numbers of singletons and doubletons increased slightly when individual population groups were considered separately.

For the west-African samples the number of singletons was one, with four doubletons, whereas for east-Africans there were three singletons and two doubletons. As for the samples grouped according to expression of a full-length (23238C) allele, or truncated (23238T) allele, there were two singletons and five doubletons for the 23238C alleles and three singletons and three doubletons for the 23238T alleles. The number of singletons and doubletons in each population was tested for deviation from a neutral model of evolution, using Fu and Li`s *D** and *F** test statistics (see Section 3.2.9.6).

The *FMO2* chimp sequence was used as a reference to infer the ancestral state of each of the human diallelic SNPs (Table 14). In almost all cases, the more common human variant corresponded to the inferred chimp allele, except for g.23238C>T (Q472X), in exon 9.

There were five nonsynonymous SNPs [other than g.23238C>T (Q472X)] that occurred at relatively high frequency in the combined African sample, g.13693T>C (F182S) (0.09) and g.13732C>T (S195L) (11%) in exon 5, g.18237G>A (R238Q) (0.09) in exon 6, g.19910G>C (R391T) (13%) in exon 7 and g.22060T>G (N413K) (0.17) in exon 8. Haplotypes had to be inferred for further analysis.

Table 14. SNP alleles and their frequencies

DbSNP # Position[a]	Amino acid change[b]	Number and frequency of derived alleles (Combined samples) n=96	Number and frequency of derived alleles (West Africa) n=48	Number and frequency of derived alleles (East Africa) n=48	Number and frequency of derived alleles on a 23238C background n=48	Number and frequency of derived alleles on a 23238T background n=48
rs2020870 (g.107A>G)	(D36G)	6 (0.06)	2 (0.04)	4 (0.08)	4 (0.08)	2 (0.04)
rs6657314 (+251T>C)		7 (0.07)	3 (0.06)	4 (0.08)	2 (0.04)	5 (0.10)
rs28369800 (+409C>T)		13 (0.14)	7 (0.15)	6 (0.13)	8 (0.17)	5 (0.10)
rs28369551 (+7586C>T)		11 (0.11)	4 (0.08)	7 (0.15)	6 (0.13)	5 (0.10)
rs28745274 (g.7695T>A)	(F69Y)	2 (0.02)	2 (0.04)	0 (0.00)	2 (0.04)	0 (0.00)
rs2020860 (g.7731T>C)	(F81S)	12 (0.13)	5 (0.10)	7 (0.15)	6 (0.13)	6 (0.13)
rs16864165 (+7883T>A)		13 (0.14)	6 (0.13)	7 (0.15)	7 (0.15)	6 (0.13)
rs28369860 (g.10951delG)	(V113fsX)	2 (0.02)	0 (0.00)	2 (0.04)	2 (0.04)	0 (0.00)
rs2307492 (g.13693T>C)	(F182S)	9 (0.09)	2 (0.04)	7 (0.15)	5 (0.10)	4 (0.08)
rs2020862 (g.13732C>T)	(S195L)	11 (0.11)	3 (0.06)	8 (0.17)	4 (0.08)	7 (0.15)
rs2020861 (+13733G>A)	(S195S)	53 (0.55)	30 (0.63)	23 (0.48)	25 (0.52)	28 (0.58)
rs7542361 (+13841C>T)		8 (0.08)	4 (0.08)	4 (0.08)	5 (0.10)	3 (0.06)
rs2075987 (+13952C>A)		53 (0.55)	30 (0.62)	23 (0.48)	26 (0.54)	27 (0.56)
rs10912559 (+17928C>T)		47 (0.49)	28 (0.58)	19 (0.40)	26 (0.54)	21 (0.44)
rs28369895 (g.18237G>A)	(R238Q)	9 (0.09)	6 (0.13)	3 (0.06)	6 (0.13)	3 (0.06)
rs2020866 (g.18269C>T)	(R249X)	1 (0.01)	0 (0.00)	1 (0.02)	0 (0.00)	1 (0.02)
rs7517460 (+18390T>C)		45 (0.47)	28 (0.58)	17 (0.35)	22 (0.46)	23 (0.48)
rs2075988 (+18559T>C)		2 (0.02)	2 (0.04)	0 (0.00)	1 (0.02)	1 (0.02)
rs2020863 (g.19679A>G)	(E314G)	4 (0.04)	3 (0.06)	1 (0.02)	2 (0.04)	2 (0.04)

rs7536646 (g.19839A>G)	(A367A)	49 (0.51)	20 (0.42)	29 (0.60)	22 (0.46)	27 (0.56)
rs28369899 (g.19910G>C)	(R391T)	12 (0.13)	6 (0.13)	6 (0.13)	7 (0.15)	5 (0.10)
rs7536745 (+19969G>A)		41 (0.43)	26 (0.54)	15 (0.31)	22 (0.46)	19 (0.40)
rs6671692 (g.22027G>A)	(E402E)	45 (0.47)	28 (0.58)	17 (0.35)	24 (0.50)	21 (0.44)
rs2020865 (g.22060T>G)	(N413K)	16 (0.17)	6 (0.13)	10 (0.21)	8 (0.17)	8 (0.17)
rs28369908 (+22216G>A)		14 (0.15)	7 (0.15)	7 (0.15)	8 (0.17)	6 (0.13)
rs2075989 (+22267T>C)		3 (0.03)	1 (0.02)	2 (0.04)	1 (0.02)	2 (0.04)
rs2075990 (+22353T>G)		9 (0.09)	4 (0.08)	5 (0.10)	2 (0.04)	7 (0.15)
rs28369911 (+23006G>T)		83 (0.86)	41 (0.85)	42 (0.88)	40 (0.83)	43 (0.90)
rs28369912 (g.23087A>G)	(G421G)	6 (0.06)	3 (0.06)	3 (0.06)	2 (0.04)	4 (0.08)
rs61730973 (g.23115A>C)	(N431T)	1 (0.01)	0 (0.00)	1 (0.02)	0 (0.00)	1 (0.02)
rs6661174 (g.23238C>T)	(Q472X)	48 (0.50)	24 (0.50)	24 (0.50)	0 (0.00)	48 (1.00)
rs2020869 (g.23300A>G)	(K492K)	11 (0.11)	4 (0.08)	7 (0.15)	5 (0.10)	6 (0.13)

n Number of alleles,

a Positions of SNPs are given relative to the A of the ATG translational initiation codon. In addition, for each of the 18 coding-region SNPs, the position and identity of the resultant change in amino acid residue is given.

b Amino acid change and the protein position at which it occurs. The ancestral state of each SNP was inferred from the chimp sequence.

SNPs, single-nucleotide polymorphisms.

3.2.3 Haplotypes

Haplotypes were inferred from genotypes using the Phase algorithm of DnaSP, version 5, and the EM and ELB algorithms of Arlequin. Both EM and Phase were more accurate than ELB when handling rare haplotypes (Sabbagh and Darlu 2005). The Phase approach was chosen because of the occurrence of rare haplotypes in the dataset and it being better in inferring the former, and also because Phase had the lowest error rate according to (Sabbagh and Darlu 2005). This is consistent with studies that evaluated the performance of various algorithms using both simulated and empirical datasets (Stephens *et al*. 2001; Stephens and Donnelly 2003). Because some analysis, such as nucleotide diversity, requires complete sequence information only individuals with complete genotypes were used. It is important to analyse a sample that represents the frequency of both 23238C and 23238T alleles in sub-Saharan Africa. Individuals have been chosen according to that criterion. In contrast, when looked at separately, the maximum number of successfully phased 23238C and 23238T alleles can be used for analysis.

3.2.3.1 Combined African sample

In the sample of 30 chromosomes, (twenty-four 23238T and six 23238C alleles), the SNPs segregate as ten distinct haplotypes (Table 15). The ten haplotypes are shown, together with the inferred ancestral state (chimp) of each allele, in (Table 16). No haplotype in the study exactly matched the ancestral haplotype. The closest were haplotypes 2, 4 and 10. The most common haplotype is haplotype 5, with a frequency of 0.40.

3.2.3.2 West- and east African samples

Eight distinct haplotypes were identified in the sample of 22 west-African chromosomes, and four in the east-African sample of eight chromosomes (Table 15).

3.2.3.3 23238C and 23238T allele groups

In the sample of twenty-four 23238C alleles, there were fourteen haplotypes, whereas the twenty-four 23238T alleles segregated as six haplotypes (Table 15).

Table 15. Number of haplotypes inferred for the combined African resequencing and for the separated groups

Population Group	**Number of haplotypes inferred**
Combined resequencing group	**10**
West-African group	**8**
East-African group	**4**
23238C allele group	**14**
23238T allele group	**6**

Table 16. Inferred haplotypes for the combined African group

Nucleotide position **Haplotype id and occurence**		+251T>C	+409C>T	+7586C>T	**+7731T>C**	+7883T>A	**+10951delG**	**+13732C>T**	+13733G>A	+13952C>A	+17928C>T	**+18237G>A**	+18390T>C	+19839A>G	**+19910G>C**	+19969G>A	+22027G>A	**+22060T>G**	+22216G>A	+22353T>G	+23006G>T	**+23238C>T**	**Frequency**
Ancestral	**0**	T	C	C	**T**	T	**G**	**C**	G	C	C	**G**	T	A	**G**	G	G	**T**	G	T	G	**C**	**0**
1	**1**	.	.	.	**.**	.	**.**	**.**	A	A	T	**A**	C	.	**.**	A	A	**.**	.	.	T	**.**	**0.033**
2	**2**	.	.	T	**C**	A	**del**	**T**	.	.	.	**.**	.	G	**.**	.	.	**.**	.	.	T	**.**	**0.066**
3	**1**	.	T	.	**.**	.	**.**	**.**	A	A	T	**.**	C	.	**C**	A	A	**.**	A	.	.	**.**	**0.033**
4	**2**	.	.	.	**.**	.	**.**	**.**	A	A	T	**.**	C	.	**.**	A	A	**.**	.	.	T	**.**	**0.066**
5	**12**	.	.	.	**.**	.	**.**	**.**	A	A	T	**.**	C	.	**.**	A	A	**.**	.	.	T	**T**	**0.40**
6	**2**	.	.	.	**.**	.	**.**	**.**	A	A	T	**A**	C	.	**.**	A	A	**.**	.	.	T	**T**	**0.066**
7	**2**	.	T	.	**.**	.	**.**	**.**	A	A	T	**.**	C	.	**C**	A	A	**.**	A	.	.	**T**	**0.066**
8	**2**	.	T	.	**.**	.	**.**	**.**	A	A	T	**.**	C	.	**C**	.	A	**.**	A	.	.	**T**	**0.066**
9	**3**	.	.	T	**C**	A	**del**	**T**	.	.	.	**.**	.	G	**.**	.	.	**G**	.	.	T	**T**	**0.10**
10	**3**	C	.	.	**.**	.	**.**	**.**	A	A	.	**.**	.	G	**.**	.	.	**.**	.	G	T	**T**	**0.10**

Haplotypes and their estimated frequencies in the combined African groups (30 chromosomes). The nucleotide positions of SNPs are given relative to the A of the ATG translational initiation codon. Those that result in amino-acid substitutions, 7731T>C (F81S), 10951delG (V113fsx), 13732C>T (S195L), 18237G>A (R238Q), 19910G>C (R391T), 22060T>G (N413K) and 23238C>T (Q472X) are shown in bold. All polymorphisms are shown as derived changes in comparison with the chimpanzee sequence. Bases identical to the chimpanzee sequence are indicated by a dot. Haplotypes are numbered arbitrarily.

3.2.4 Frequency of functional 23238C alleles in sub-Saharan Africa

To estimate the frequency of functional alleles in the study sample, the 23238C alleles from west and east Africa were pooled together, 48 chromosomes, of which 28 were phased successfully for all nonsynonymous SNPs (Table 17). The resulting eight haplotypes were examined for the presence of nonsynonymous SNPs. Results show that more than half of full-length 23238C alleles (0.54) are not associated with nonsynonymous SNPs and, therefore, are apparently functional. A reduced-median network diagram visualizes the haplotypes (Figure 14).

Table 17. Frequency of functional and non-functional 23238C alleles

Nucleotide change[a]		g.107 A>G	g.7695 T>A	g.13693 T>C	g.13732 C>T	g.18237 G>A	g.19679 A>G	g.19910 G>C	g.22060 T>G	Number and frequency	Predicted effect on FMO2 function[b]
Location		Exon 2	Exon 3	Exon 5	Exon 5	Exon 6	Exon 7	Exon 7	Exon 8		
Amino acid change		D36G	F69Y	F182S	S195L	R238Q	E314G	R391T	N413K		
Haplotype id	Ancestral	A	T	T	C	G	A	G	T	15 (0.54)	Functional
	1	G	A	C	T	G	A	G	T	1 (0.035)	Non-functional
	2	G	T	T	C	G	A	G	T	1 (0.035)	Non-functional
	3	G	T	T	C	G	G	G	T	1 (0.035)	Non-functional
	4	A	T	C	C	G	A	G	T	2 (0.07)	Non-functional
	5	A	T	T	C	A	A	G	T	4 (0.14)	Non-functional
	6	A	T	T	C	G	A	C	T	3 (0.11)	Non-functional
	7	A	T	T	C	G	A	G	G	1 (0.035)	Probably functional

[a]Position from the A of the ATG initiation codon in the *FMO2* genomic reference sequence (AL021026).
Total 23238C alleles=28.
Nucleotide changes that are predicted to cause loss of function are underlined.
Haplotype id refers to the numbering of the haplotypes in the RM network diagram (Figure 14).
[b]Effect on function was according to prediction by PolyPhen based on single amino-acid changes.

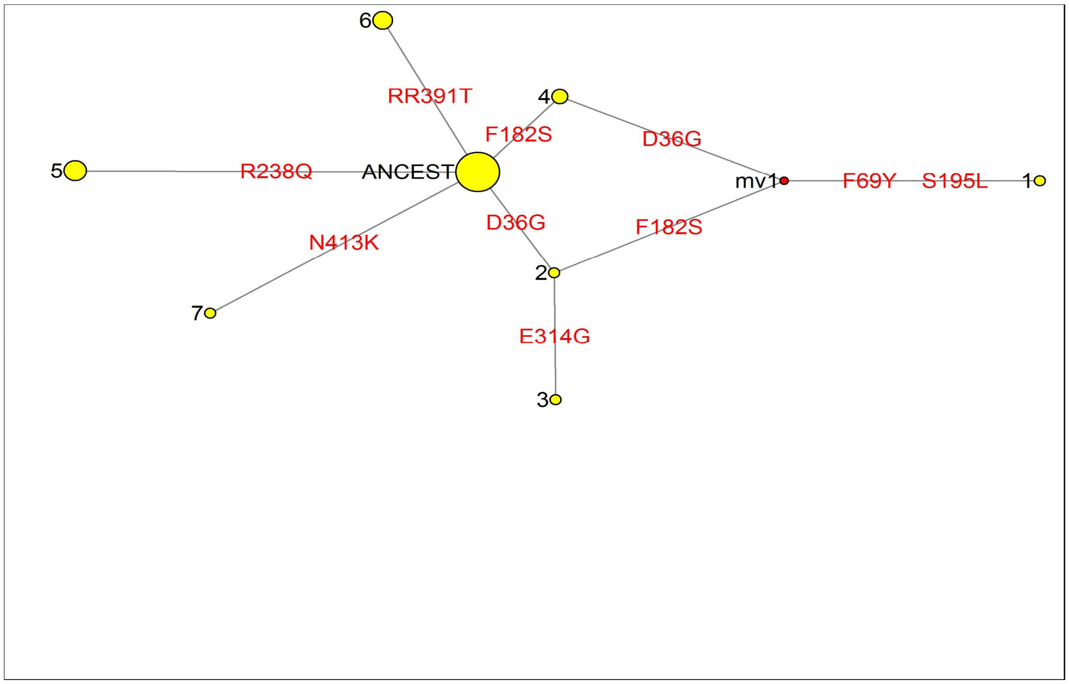


Figure 14. Reduced-median network of *FMO2* haplotypes for the 23238C allele group of 28 chromosomes. Each of the eight unique haplotypes is represented by a circle, the size of which is proportional to the relative frequency of the haplotype. Mutational relationships are indicated by lines linking the haplotypes. Mutational differences between haplotypes are indicated on the branches of the network. Mutations are identified by the position of the amino-acid residue in the protein. Where multiple mutations are present on the same line, the order in which they occurred is not known.
A, Ancestral.

3.2.5 Summary of the variation observed in sub-Saharan African populations and in the NIEHS African population

A total of 32 variants were observed in sub-Saharan populations, whereas 24 were documented in the corresponding *FMO2* sequence in the African NIEHS population. The ratio of coding to non-coding variants was 18:14 in Africans and 16:8 in the NIEHS African population (Fisher`s exact test, *P>0.3*). It was 15:14 in 23238C alleles and 15:14 in the 23238T alleles. West-Africa had a ratio of 15:14, whereas east-Africa had a ratio of 18:13 (Fisher`s exact test, *P*>0.4).

The transition to transversion ratio was 23:8 in Africans, compared with 16:4 in the NIEHS population (Fisher`s exact test, *P>0.4*), 21:7 in the 23238C alleles and 22:7 in the 23238T alleles (Fisher`s exact test, *P*>0.2), 22:7 for west-Africa and 23:8 in east-Africa (Fisher`s exact test, *P*>0.5)

The ratio of synonymous to nonsynonymous variants was 5:12 in Africans and 2:11 in NIEHS (Fisher`s exact test, *P>0.3*). It was 5:11 in the 23238C alleles and 5:10 in the 23238T alleles (Fisher`s exact test, *P>0.2*). It was 5:9 in west-Africa and 5:9 in east-Africa.

3.2.6 Prediction of effect of nonsynonymous SNPs on function

PolyPhen software, www.genetics.bwh.harvard.edu/pph/, was used to predict the effect on protein function of each of the 18 c.SNPs in this study (Table 18). In addition to the SNPs at position g.10951delG (V113fsX), g.13732C>T (S195L) (Furnes *et al*. 2003; Krueger 2005) and g.23238C>T (Q472X) (Dolphin 1998; Whetstine *et al*. 2000), which are known individually to cause loss of function, other SNPs that are predicted to be damaging by PolyPhen are: g.107A>G (D36G), g.7731T>C (F81S), g.13693T>C (F182S), g.18237G>A (R238Q), g.18269C>T) (R249X), g.19679A>G (E314G), g.19910G>C (R391T) and g.23115A>C (N431T). The remaining SNPs are predicted to be benign.

3.2.7 Recombination

The presence of pairs of sites with four-gametic combinations is indicative of recombination (crossing-over between the two sites or gene conversion), or repeated mutation. Repeated mutation at the same site has a very low probability. Consequently, out of the 231 SNP pairs compared by the four-gamete test (Hudson and Kaplan 1985), the presence of the 35 four-gamete site pairs is most likely accounted for by recombination. The four-gamete test predicts a minimum of five recombination events in the combined African group: between (+409, g.13733), (g.18237, g.23238), (+13952, +18390) and (g.19910, +19969).

Table 18. PolyPhen prediction of the effect of c.SNPs on FMO2 function

SNP	Position in Gene	Amino acid Position	PSIC score difference	Prediction	SNP category	Reference
rs2020870	Exon 2	D36G	2.425	Probably damaging	Missense (radical)	n.d
rs28745274	Exon 3	F69Y	0.058	Benign	Missense (conservative)	n.d
rs2020860	Exon 3	F81S	2.771	Probably damaging	Missense (radical)	n.d
rs28369860*	Exon 4	V113fsX	1.568	Possibly damaging	Frameshift	(Furnes *et al*. 2003; Krueger *et al*. 2005)
rs2307492	Exon 5	F182S	2.775	Probably damaging	Missense (radical)	
rs2020862*	Exon 5	S195L	2.702	Probably damaging	Missense (radical)	(Furnes *et al*. 2003; Krueger *et al*. 2005)
rs28369895	Exon 6	R238Q	2.314	Probably damaging	Missense (radical)	

rs2020866	**Exon 6**	**R249X**	**2.141**	**Probably damaging**	**Nonsense (stop-codon)**	
rs2020863	**Exon 7**	**E314G**	**2.190**	**Probably damaging**	**Missense (radical)**	
rs28369899	**Exon 7**	**R391T**	**2.355**	**Probably damaging**	**Missense (radical)**	
rs2020865*	**Exon 8**	**N413K**	**0.645**	**Benign**	**Missense (conservative)**	**(Furnes *et al*. 2003; Krueger *et al*. 2005)**
rs61730973	**Exon 9**	**N431T**	**0.649**	**Benign**	**Missense (conservative)**	
rs6661174*	**Exon 9**	**Q472X**	**1.988**	**Possibly damaging**	**Nonsense (stop-codon)**	**(Dolphin *et al*. 1998; Whetstine *et al*. 2000)**

Table displays the effect of c.SNPs on protein function, predicted by PolyPhen, the accession number of the SNP, its location in the gene, the amino-acid change and the SNP category. *Indicates SNPs the function of which have been determined experimentally. The last column shows references reporting the functional effect of these SNPs.

3.2.8 Linkage disequilibrium

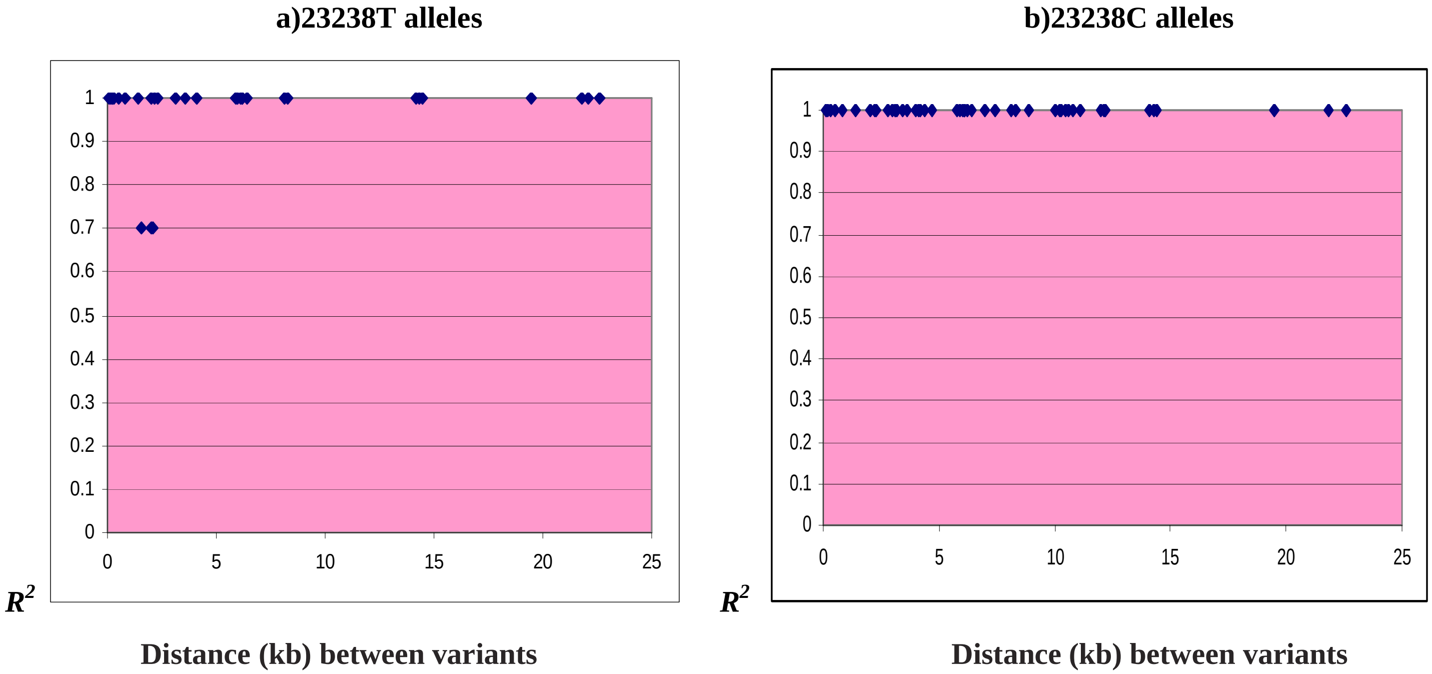
Linkage disequilibrium (LD) was examined for the combined African group initially (results not shown) and then for individual groups. Interesting results were found when grouped into 23238T alleles. A plot of pairwise LD (measured as R^2) against physical distance for the samples divided into truncated (Ts), indicates that relatively strong LD extends for a distance of 23 kb (Figure 15a), between the SNP at +409 and that at +23006, which lies some 22.6 kb apart (Figure 16).

For the 23238C alleles, none of the ten pairwise comparisons between the five SNPs show significant LD, whereas for the 23238T alleles, 32 of the 153 comparisons show significant LD (*P<0.001*) (Figure 16). The absence of any significant pairwise comparisons for the 23238C alleles may be due to the small sample number (six alleles). To test this, and since the 23238C alleles will be looked at separately and not in the context of a combined sample, which has to reflect the 23238C:23238T allele ratio found in sub-Saharan Africa, all 48 23238C alleles were pooled together and 24 were successfully phased for SNPs. Pairwise LD comparisons were performed and results show that of the 120 pairwise comparisons between the 14 SNPs, 51 show significant LD with Bonferroni correction (*P*<0.001) (Figure 15b), and, as is the case for the 23238T alleles, LD extends for 22.6 kb (Figure 17)

To assess the level of LD of *FMO2*, five test statistics, Z_{nS} (Kelly 1997), Z_a (Rozas *et al*. 2001), ZZ (Rozas *et al*. 2001) Wall's *B* and Wall's *Q* (Wall 1999), were performed (Tables 19 and 20). No significant values were found for any of the five test statistics for either the 23238C or 23238T alleles (Table 20), whereas significant values of Z_a and Z_{nS} were observed for the NIEHS Asian group

(*P*<0.05) (Table 19). Z_a had a higher than expected value, whereas Z_{nS} had a lower than expected value under neutrality. Significant LD for 11 SNPs in the NIEHS Asian group extends for 22.2kb across the *FMO2* gene, between the SNPs at +107 and +22267, for example (results not shown). Twenty-eight of the 66 pairwise comparisons are significantly different with Bonferroni correction (*P*<0.001) (Figure 18). The value of Z_{nS} may suggest that the *FMO2* gene in Asians has been subjected to some form of selection.

Significantly higher than expected positive values for ZZ were observed for the NIEHS 23238T allele group (consists of 23238T alleles from African, European, Hispanic and Asian individuals) (*P*<0.001), this suggests intragenic recombination. Significant LD for 19 SNPs extends for 23.1kb across the *FMO2* gene, for example, between +251 and +23353. Forty-six of the 190 pairwise comparisons show significant LD with Bonferroni correction (*P*<0.001) (results not shown). All statistical tests performed by DnaSP were tested by 5000 coalescent simulations in the 95% confidence interval. Significant LD was also observed in the NIEHS European group. Significant LD for 10 SNPs extends for 22.1kb for example between +251 and +22353 (results not shown). Twenty-one of the 45 pairwise comparisons show significant LD with Bonferroni correction (*P*<0.001) (Figure 19).



Figures 15a and b. Plot of R^2 against distance.

15a) Relationship between pairwise LD measured as R^2 and physical distance (in kb) for the 32 associations that display significant LD at the 0.001% level in 23238T alleles.

15b) Relationship between pairwise LD measured as R^2 and physical distance (in kb) for the 51 associations that display significant LD at the 0.001% level in 23238C alleles.

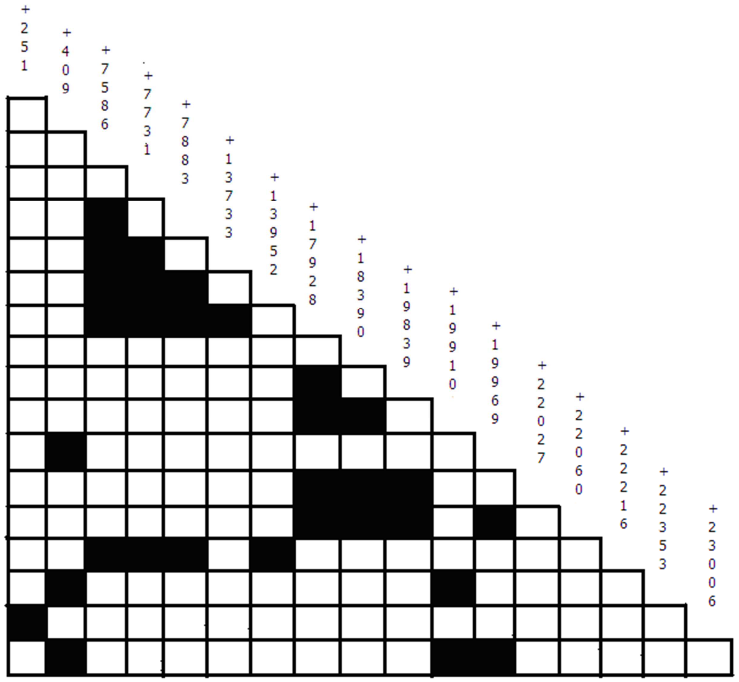


Figure 16. Plot of the association among 17 SNPs for 12 African 23238T homozygotous individuals.

Blackened boxes indicate the 32 out of 153 possible associations that display significant LD at the 0.001% level by a x_2 test with Bonferroni correction. Numbers denote the nucleotide positions of the SNPs relative to the A of the ATG translational initiation codon.

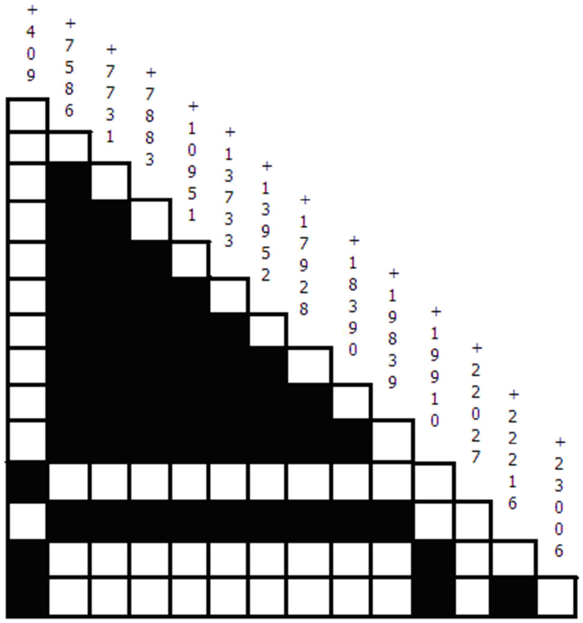


Figure 17. Plot of the association among 14 variants for 12 African 23238C homozygotous individuals.

Blackened boxes indicate the 51 out of 120 possible associations that display significant LD at the 0.001% level by a x_2 test with Bonferroni correction. Numbers denote the nucleotide positions of the SNPs relative to the A of the ATG translational initiation codon. All are SNPs except for one indel:g.10951delG

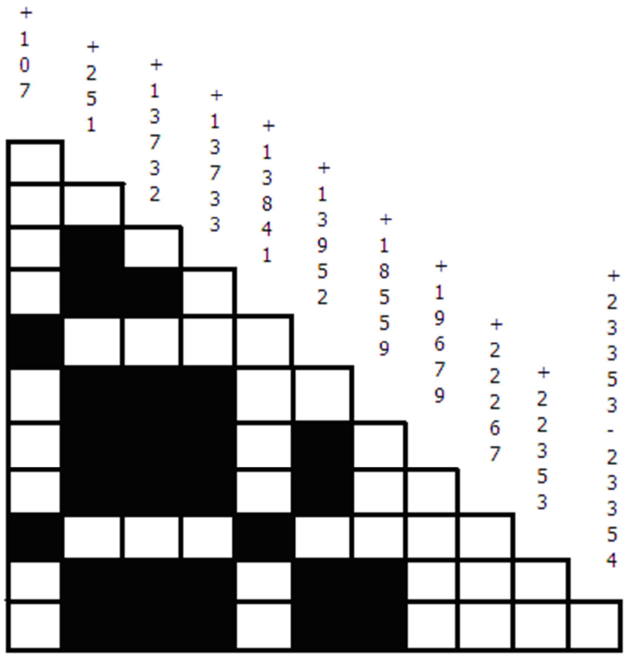


Figure 18. Plot of the association among 11 variants for 22 NIEHS Asian 23238T homozygotous individuals. Blackened boxes indicate the 28 out of 66 possible associations that display significant LD at the 0.001% level by a x_2 test with Bonferroni correction. Numbers denote the nucleotide positions of the SNPs relative to the A of the ATG translational initiation codon. All are SNPs except for two indels: g.13592delA and g.23353-23354insT. Positions of the SNPs relative to the A of the ATG translational initiation codon.

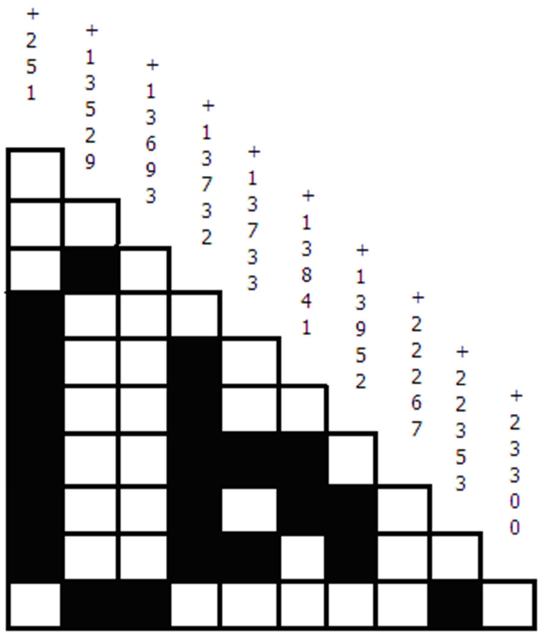


Figure 19. Plot of the association among 10 variants for 14 NIEHS European 23238T homozygotous individuals. Blackened boxes indicate the 12 out of 45 possible associations that display significant LD at the 0.001% level by a x_2 test with Bonferroni correction. Numbers denote the nucleotide positions of the SNPs relative to the A of the ATG translational initiation codon. All are SNPs except for one indel: g.23353-23354insT.

- It has to be noted that when 23238C and 23238T allele groups (both from the combined African resequencing samples and the NIEHS samples) were analyzed separately for various neutrality tests (Tables 19 and 20), the assumptions of these tests were opposed. The analysis of these separated groups was still performed in order to get insights into what is occurring in these groups.

3.2.9 Further testing for selection

3.2.9.1 Nucleotide diversity π and θ

Several tests were used to investigate the compatibility of the observed variation in the sequenced region with neutral evolution. These were based on (i) the allele frequency spectrum within a locus, (ii) the number and diversity of haplotypes in a sample, (iii) LD and (iv) interspecific comparisons of sequence variation. The two estimates of nucleotide diversity (theta estimates), θ_S and π, should be equal under neutrality, random mating and a constant population size, and, therefore, Tajima`s *D* statistic, which compares these two estimates, should be zero. Calculations of measures of nucleotide diversity were based on the number of effectively silent sites in the sequenced region. This was estimated to be 3442 and is the sum of all intronic and silent exon sites. The average expected per-site nucleotide heterozygosity (θ_S), estimated from the observed number of polymorphic sites (Sachidanandam *et al*. 2001), and nucleotide diversity (π), a direct estimate of per-site heterozygosity, derived from the average pairwise sequence difference between two random sequences in a sample (Bamshad and Wooding 2003), were calculated.

None of the observed values for π or θ_S were significantly different from what is expected under neutrality in any of the groups. The 23238C allele group had higher nucleotide diversity (π) compared with the 23238T allele group: 0.00143 and 0.00110, respectively (Table 20). East Africa had higher π compared with west Africa: 0.0126 and 0.00116, respectively (Table 19). In the NIEHS groups, the highest π was observed in both the NIEHS African and Asian groups: 0.0015 and 0.0015, respectively, this was followed by the NIEHS European group, 0.0012, and the NIEHS Hispanic group, 0.0011 (Table 19). For the combined sequenced African samples, θ_S= 0.00118 and π = 0.00120 (Table 19). In this study the estimates of nucleotide sequence diversity (π) for *FMO2* are higher than the average value of π (0.00075) for human genetic loci (Sachidanandam *et al*. 2001; Bamshad and Wooding 2003). It was found that for the NIEHS samples, both estimates, θ_S = 0.0012 and π = 0.0013, are lower than those observed for the combined sequenced African samples, but still higher than the average for human loci. All the observed π values are in the range mentioned by Sachidanandam *et al*. (Sachidanandam *et al*. 2001), which was 2-15.8 x 10^{-4}.

3.2.9.2 Tajima`s *D* statistic

A significantly positive Tajima`s *D* statistic was observed for the NIEHS Asian group (P<0.001) (Table 19). This implies an excess of moderate-frequency, compared with low-frequency, variants. This may be due to population structure, another possibility is balancing selection, which can also lead to an excess of moderate-frequency variants, since in balancing selection more than one allele is maintained at moderate frequency in a population.

Table 19. Diversity estimates and neutrality tests for FMO2

Sample Summaries	NIEHS	African	European	Hispanic	Asian	NIEHS 23238Cs	NIEHS 23238Ts	Combined	W.A	E.A
n^a	106	10	28	24	44	6	100	30	22	8
S^b	31	17	18	17	15	13	25	22	19	15
s^c	8	4	4	4	4	1	2	2	0	2
h^d	29*	8	12	12	10	6	23***	10	8	4
h^e	19	6	10	15	10	5	16	11	9	6
$R_m{}^f$	6*	1*	5*	3*	1*	2*	4*	5*	1*	2*
Hd^g	0.824	6%95	0.812	0.844	0.866	1	0.801	0.825	0.840	0.750
π^h	0.0012	0.0015	0.0012	0.0011	0.0015	0.0015	0.0010	0.00120	0.00116	0.00126
$\theta_S{}^i$	0.0013	0.0013	0.001	0.0009	0.0007	0.0012	0.00103	0.00118	0.00111	0.00123
Tajima`s *D*	-0.12197	0.76663	0.6761	-0.9939	3.3334#	1.47467	0.20838	0.01519	0.1689	0.19202
Fu and Li`s *D**	-0.66596	0.61631	0.26441	0.2467*	-0.2091	0.93041	1.09310*	0.79056	1.5811*	0.37963
Fu and Li`s *F**	-0.54153	0.73656	0.46136	0.40154	1.14675	1.13464	0.90215	0.65784	1.350*	0.35618
Fu and Li`s *D*	1.02540	0.83587	-3.101#	0.11403	1.66247*	-0.68740	1.91496#	0.84864	1.7839*	0.24594
Fu and Li`s *F*	0.59620	1.04734	-2.81041*	0.28716	2.58096*	-0.38230	1.65970*	0.70170	1.5235*	0.25994
Fu`s *F*s	-6.833	-1.127	-0.8070	-1.582	3.372**	-1.521	-3.545	0.206	1.356	2.645
Fay and Wu's *H*	-2.47547	1.24444	-10.1935	2.41304	-15.8118#	-0.6060	-11.2460*	-3.779	-4.31169	-2.14286
Raggedness	0.0401	0.0583	0.0506	0.0572	0.1046**	0.0800	0.0630	0.0287	0.0501	0.2691
Z_{nS}	0.1277	0.3135	0.1972	0.1727	0.3446**	0.4185	0.1233	0.2046	0.2964	0.3760
Z_a	0.2093	0.4001	0.2255	0.2485	0.5104**	0.5727	0.3154	0.2140	0.3999	0.3211
ZZ	0.0816	0.0866	0.0282	0.0758	0.1658	0.1542	0.1922#	0.0093	0.1035	-0.0549
Wall`s *B*	0.1500	0.2667	0.0000	0.1333	0.0000	0.4545	0.2500	0.0476	0.3333	0.0714
Wall`s *Q*	0.2381	0.3750	0.0000	0.2500	0.0000	0.5833	0.4286	0.0909	0.4211	0.1333*
R_2	0.0907	0.1865	0.1475	0.1448	0.2328*	0.2278	0.1027	0.1265	0.1434	0.1790

a Number of chromosomes surveyed.
b Number of segregating sites.
c Number of singleton sites.
d Number of observed haplotypes.
e Number of expected haplotypes from coalescent simulations.
f Minimum number of recombination events, based on the four-gamete test.
g Haplotype diversity.
h Average pairwise sequence difference per nucleotide assuming no recombination.
R_2 Ramos-Onsins and Rozas.

A total of 4705 sites (3442 silent sites) were analysed, *P* Values are given only when significant *(*P*≤0.05), no recombination, **(*P*≤0.05), with recombination , ***(*P*≤0.05), without recombination, #(*P*≤0.001) no recombination, based on estimates from coalescent simulations E.A, East Africa, W.A, West Africa., combined, combined African resequencing samples.

Table 20. Diversity estimates and neutrality tests for FMO2 for the 23238C and T alleles

Sample Summaries	23238Cs	23238Ts
n[a]	24	24
S[b]	23	18
s[c]	1	0
h[d]	14	6
h[e]	10	9
R_m[f]	3*	2*
Hd[g]	0.917	0.728
π[h]	0.00143	0.00110
θ_S[i]	0.00131	0.00102
Tajima\`s *D*	0.34780	0.35824
Fu and Li\`s *D**	0.36284	1.57669*
Fu and Li\`s *F**	0.41904	1.37593*
Fu and Li\`s *D*	0.34900	1.76324*
Fu and Li\`s *F*	0.42290	1.53674*
Fu\`s *F*s	-2.230	3.47*
Fay and Wu's *H*	1.62319	-2.47101
Raggedness	0.0179	0.1127
Z_{nS}	0.1726	0.3225
Z_a	0.2152	0.4011
ZZ	0.0427	0.0787
Wall\`s *B*	0.0909	0.2941
Wall\`s *Q*	0.1739	0.3889
R_2	0.1403	0.1437

a Number of chromosomes surveyed.
b Number of segregating sites.
c Number of singleton sites.
d Number of observed haplotypes
e Number of expected haplotypes from coalescent simulations.
f Minimum number of recombination events, based on the four-gamete test.
g Haplotype diversity.
h Average pairwise sequence difference per nucleotide assuming no recombination.
I Expected heterozygosity per nucleotide assuming no recombination.
R_2 Ramos-Onsins and Rozas. Number of expected *P* Values are given only when significant.
A total of 4705 sites (3442 silent sites) were analysed,*($P \leq 0.05$), no recombination, **($P \leq 0.05$), with recombination, ***($P \leq 0.001$) no recombination, based on estimates from coalescent simulations.

3.2.9.3 Haplotype diversity and Fu`s *F*s statistic

Haplotype diversity is a measure of the uniqueness of haplotypes in a given population, whereas Fu`s *F*s statistic tests whether the number of haplotypes observed in a population is similar to that expected under neutrality. Haplotype diversity for all groups was not different from what is expected under a neutral model of evolution (Tables 19 and 20).

There was an excess of haplotypes in two groups. The excess was noticed in the combined NIEHS group and in the separated NIEHS 23238T allele group (Table 19). For example, the NIEHS group had 29 haplotypes instead of the expected 19. The observed numbers were significantly different from the expected numbers under a neutral model for the two groups mentioned above ($P<0.05$).

3.2.9.4 Pairwise number of differences and raggedness (r)

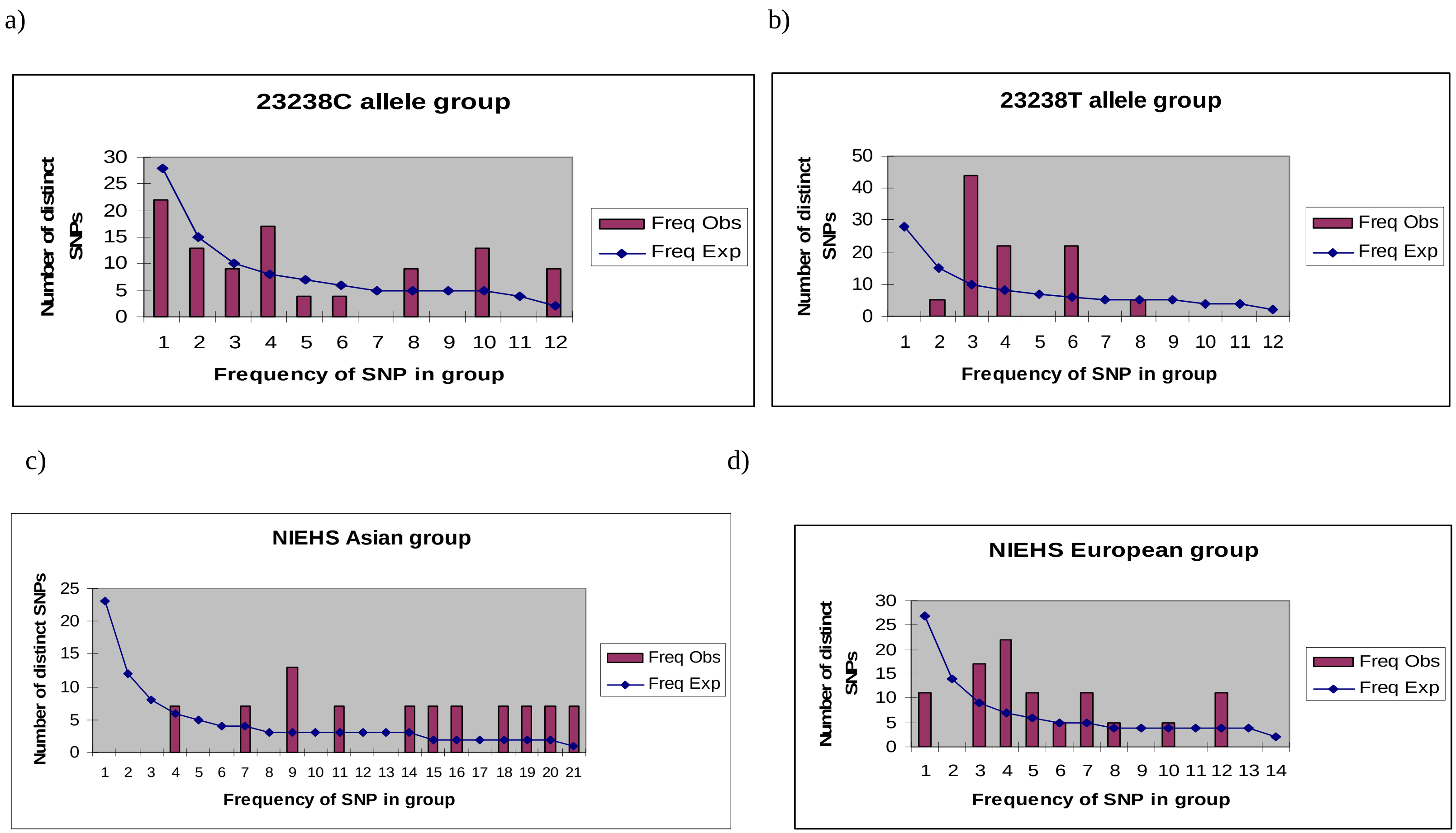
The pairwise number of differences, also called Roger and Harpingding`s mismatch distribution, was estimated for all population groups. This test shows the observed pairwise nucleotide differences and the expected values in a population with constant size, without recombination, as well as with change in population size (growth-decline) with free recombination (Rogers and Harpending 1992) (results not shown). The average number of pairwise differences for the combined NIEHS group was 5.7, for the NIEHS Asian, European and Hispanic group it was 7.1, 5.5 and 9.7, respectively, whereas for the west-African group it was 5.4. The 23238C and 23238T allele groups had an average number of 6.9 and 5.5, respectively, whereas the NIEHS 23238C and 23238T alleles group had an average of 7.1 and 5.2, respectively.

The raggedness statistic is shown in Tables 19 and 20 and is used to quantify the smoothness of the observed pairwise differences distribution. The raggedness value (r) is significantly different from that expected in the NIEHS Asian group (0.1046) (*P*<0.05). The (r) value of raggedness for the NIEHS Asian group is either indicative of a population size that has been constant for a long time or balancing selection (Rogers and Harpending 1992). Caution has to be exercised since the raggedness statistic has low statistical power for detecting population expansion; a more powerful test is Fu`s *F*s (see Section 3.2.9.33.2.9.3) and the Ramos-Onsins and Rozas`s R_2 (Ramos-Onsins and Rozas 2002). The R_2 statistic is used to detect recent population growth, which is indicated by low values of the test statistic.

It is based on the difference between the number of singleton mutations and the average number of nucleotide differences. The R_2 value for the NIEHS Asian group was higher than expected under neutrality and not low, so no recent population expansion was detected, therefore the observed increase in low-frequency alleles for the NIEHS Asian group (see Section 3.2.9.7), can not be explained by population expansion. A plausible explanation for all these observations in addition to the excess in intermediate-frequency variants (see below for allele-frequency spectrum) could be a bottleneck after the migration of modern *Homo sapiens* out of Africa, which led to a decrease in population size and consequently, to an increase in genetic drift (Rosenberg *et al*. 2002).

3.2.9.5 Allele-frequency spectrum

The allele-frequency spectrum of *FMO2* for the 23238C and 23238T allele groups, as well as for the NIEHS Asian and European group indicates an excess of intermediate-frequency variants, when compared with a neutral model (Figures 20a, b, c and d).



Figures 20a, b, c and d. Allele frequency spectrum of derived SNP alleles. Bars indicate the observed frequencies, whereas the dotted line indicates the frequencies expected under a neutral model of neutrality. SNP, single- nucleotide polymorphism.

3.2.9.6 Fu and Li`s *D** and *F** test statistics

Fu and Li`s *D\** and *F\** test statistics (Fu and Li 1993) compare the observed number of singleton polymorphisms with the number expected under a neutral model. The values of Fu and Li`s *D** were significantly positive in the west-African group (P<0.05), the 23238T allele group (P<0.05) (Tables 19 and 20), the NIEHS Hispanic group (P<0.05) and the NIEHS 23238T allele group (P<0.001). Fu and Li`s *F** was significantly positive in the west-African (P<0.05) and 23238T allele groups (P<0.05) (Tables 19 and 20). Positive values are an indication of old versus young mutations which might be caused by balancing selection or population subdivision.

An approach to test for population subdivision, Wright's fixation index (Wright 1931), which measures population subdivision (or genetic differentiation) by comparing genetic diversity within subpopulations with that in the whole population, was used to analyse the heterozygosity of 32 *FMO2* variants. The average F_{ST} value among the 23238C alleles and 23238T alleles is 0.11, which is high. This might be due to population subdivision, which is a possibility since populations from both west and east Africa were included in each of the 23238C and 23238T allele groups. It could also be explained by the effect of a selective sweep acting at the g.23238C>T (Q472X) locus in some groups and not others and, therefore, resulting in high differentiation between populations. A complete selective sweep decreases the level of polymorphism in the selected allele due to the hitchhiking effect in which neutral and nearly neutral genetic variation linked to the selected mutation rises in frequency together with the selected mutation (Figure 21), resulting in very low genetic diversity, since the selected allele and the hitchhikers are predominant for a long time. This will be accompanied by strong LD between loci, which will need a long time to be broken down by recombination or the accumulation of new mutations. It takes approximately 2 myr to regain 0.75 of genetic diversity (Diller *et al*. 2002). Partial selective sweeps could be detected by examining intra-specific variation of putatively neutral regions (e.g., intronic sequences surrounding the g.23238C>T (Q472X) locus. This will be examined later in the thesis (see Section 3.2.9.12.1).

The average F_{ST} among the NIEHS populations (African-American, African, European, Hispanic and Asian) = 0.20, which is very high. This high value can either be indicative of high population subdivision, which can be explained by the geographical locations of the individuals included, but there also is the possibility that the high value is due to selection acting on one population, therefore resulting in high diversity. Pairwise F_{ST} for 24 variants was estimated and results show that African populations were the most diverse and furthest away from other populations (Table 21).

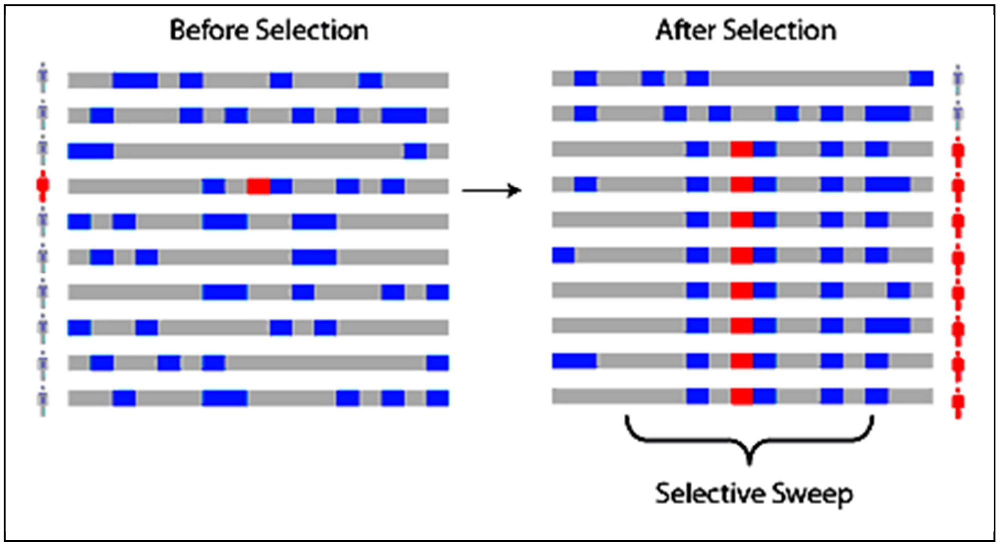


Figure 21. A selective sweep. The diagram shows the pattern of polymorphisms along a chromosome, with a selected allele, before (left) and after (right) selection. Ancestral alleles are shown in grey, derived alleles in blue, while the selected allele is in red. As the selected allele increases in frequency, closely linked mutations `hitchhike` along to high frequency, resulting in a selective sweep. Diagram taken from (Schaffner and Sabeti 2008).

Table 21. Pairwise F_{ST} comparisons among NIEHS populations

Population			
	African	**European**	**Hispanic**
European	**0.16**		
Hispanic	**0.14**	**0.00**	
Asian	**0.19**	**0.06**	**0.00**

Pairwise F_{ST} was between pairs of NIEHS populations for 24 variants. African population is made up of African-Americans and of Africans from Yoruba, Nigeria.

3.2.9.7 Fu and Li`s *D* and *F* test statistics with an outgroup

The significance of all obtained values for the individuals analysed was tested by comparison with those obtained from coalescent simulations with the same sample size as the observed data. Significantly positive values of Fu and Li`s *D* statistic (Table 19) were observed for the west-African group ($P<0.05$), the NIEHS Asian group ($P<0.05$) and the NIEHS 23238T allele group ($P<0.001$). A significantly negative value of Fu and Li`s *D* was observed for the NIEHS European group ($P<0.001$). Significantly positive values of Fu and Li`s *F* statistic were observed for the west-African ($P<0.05$), the NIEHS Asian group ($P<0.05$) and the NIEHS 23238T allele group ($P<0.05$). A significantly negative value of Fu and Li`s *F* statistic was observed for the NIEHS European group ($P<0.05$).

This value was obtained from coalescent simulations with the same sample size as the test sample with no recombination. Significantly positive values for Fu and Li`s *D* statistic indicate that the number of singletons observed in the sample exceeds the number expected under a neutral model of evolution, whereas negative values indicate fewer than expected singletons under neutrality. The excess in singletons in the west-African sample may be due to population expansion, this is supported by the low value of R_2, although not significant, whereas for the NIEHS Asian group, the positive Fu and Li`s *D* value may be due to population structure, since the value of R_2 was not low, so recent population expansion can be ruled out and does not account for the excess in singletons.

3.2.9.8 Fay and Wu`s *H*

Results of Fay and Wu`s *H* are shown in Tables 19 and 20. The Fay and Wu`s *H* statistic tests for the presence of derived alleles at a high frequency (Fay and Wu 2000). In contrast to Tajima`s *D* test, which is sensitive to population expansion (because the number of segregating sites responds faster to changes in population size than nucleotide heterozygosity), Fay and Wu`s *H* is not. So using both tests together may allow us to distinguish population expansion from natural selection.

The NIEHS Asian and the NIEHS 23238T allele groups had significantly negative *H*, -15.8118 (*P*<0.001) and -11.2460 (*P*<0.05), respectively. This value is caused by the presence of derived alleles at high frequency, which can be indicative of positive selection of a derived allele, which then rises in frequency.

The results of the neutrality tests for the NIEHS Asian group: significantly positive Tajima`s *D* (*P*<0.001) and Fu and Li`s *D* and *F* (*P*<0.05) and significantly negative Fay and Wu`s *H* test statistic (*P*<0.001) which indicates a higher than expected frequency of derived alleles, may be the result of the bottleneck that occurred after the migration out of Africa.

3.2.9.9 McDonald-Kreitman test

The results of the McDonald-Kreitman test for the combined African samples are shown in Table 22. Comparing human *FMO2* sequences with homologous sequences of chimpanzee (*Pan troglodytes*), revealed 31 fixed differences, 9 of these were in the coding regions, of which 5 were nonsynonymous changes [g.7731T>C (F81S), in exon 3; g.13693T>C (F182S), in exon 5; g.19679A>G (E314G), in exon 7, g.23115A>C (N431T) and g.23238C>T (Q472X) in exon 9].

Under neutral evolution, the ratio of replacement to silent mutations between species should be equal to the within species ratio. An advantage of this test of neutrality is that it is not dependent on assumptions about population history (i.e., is not affected by demographic processes).

The test was performed for both the combined African group (Table 22) and separated 23238C allele and the 23238T allele groups (results not shown). Results for the combined (*P*<0.03) and the 23238T allele groups (*P*<0.03) indicate an excess of intraspecies replacement polymorphisms (nonsynonymous SNPs) that is significantly different from neutrality. The 23238C alleles showed a similar result, although not significant (*P*>0.05). One explanation for the observed excess may be positive selection acting on the 23238C alleles, which results in an excess of nonsynonymous SNPs, which have reached intermediate frequency.

Table 22. McDonald-Kreitman test of Neutrality

	Silent	**Replacement**	**Fisher`s exact test**
Fixed	**31**	**4**	***P*=0.02***
Polymorphic	**13**	**9**	

Numbers of fixed differences found at both silent and replacement (i.e., those that change the amino-acid residue) sites in 30 African chromosomes are contrasted with a 2x2 Fisher`s exact test of independence.

3.2.9.10 K_a/K_s ratio

A K_a/K_s ratio of 1 may indicate neutral evolution, or a combination of purifying and positive selection at different loci in the gene, in a way that they cancel each other; a ratio of >1 implies positive Darwinian selection, whereas a ratio of <1 implies purifying selection. The K_a/K_s ratio was used to test for both purifying and positive selection in the data set divided into 23238C and 23238T allele groups. Comparisons of the human *FMO2* gene with *FMO2* from chimp and mouse were used as well. Results are shown in Table 23.

Systematic bias in the frequency at which nucleotides are changed has to be taken into consideration, since certain mutations are more probable than others (Bielawski and Yang 2000). Some lineages, for example, may more frequently change C to T than C to A. Three different programs (Datamonkey, DnaSP and the Ka/Ks calculator) were used for the estimation of the K_a/K_s ratio between human-mouse and human-chimp sequences. Table 23 shows the average of the ratios obtained from these programs for the 23238C and 23238T allele groups with mouse and chimp lineages, respectively.

Datamonkey (http://www.datamonkey.org) is a web-based, user-friendly online interface. Datamonkey aligns homologous DNA sequences in an attempt to infer signatures of natural selection according to a codon-based model of evolution. This is performed by estimation of the K_a/K_s ratio. Clustal X was used for the alignments. Datamonkey rejects any alignments containing stop-codons, therefore, all stop-codons were removed using an available function which cleans stop-codons. The mean K_a/K_s ratio for human 23238C and 23238T alleles in comparison with mouse were 0.30 and 0.35, respectively; with chimp the ratios were 1.20 and 0.40 respectively. The low values of the human-mouse for both the

23238C and 23238T allele groups could be explained by *FMO2* being conserved in many mammals (rabbit, mouse and rat for example) (Dolphin *et al*. 1992), whereas the high values of human-chimp for the 23238C allele group compared with human-mouse 23238 allele group may indicate that the human 23238C allele has been positively selected. This has been further investigated in this thesis (see Section 3.2.9.12).

Table 23. Average of the mean K_a/K_s ratios

Population group	**Human-mouse**[a] K_a/K_s	**Human-chimp**[a] K_a/K_s
23238C alleles	**0.30**	**1.20**
23238T alleles	**0.35**	**0.40**

Results of the mean K_a/K_s ratios for human-mouse and human-chimp sequences.
[a] Average for three methods

The significantly positive values of Fu and Li`s *D**, *F**, *D* and *F* test statistics for the 23238T allele and the west-African group (*P*<0.05) (Tables 19 and 20) may indicate selection acting at the human *FMO2* locus. Therefore, further analysis had to be performed to determine the type of selection. A plausible explanation could be purifying selection (in the west-African group, but not the 23238T alleles, which are null alleles), since variation in non-coding sequences is more tolerated than coding-sequence variation. Testing for purifying selection was performed by various methods (see below). Comparing nucleotide diversity (π) between the 23238C and 23238T allele groups in intronic sequences will be informative in attempting to determine whether positive selection has been acting on FMO2. Another possibility is that the observed values are a result of population subdivision, since both the west-African and the 23238T allele samples are from various population groups.

3.2.9.11 Testing for purifying selection

Purifying selection is a form of natural selection, also called negative selection. It selectively acts against mutations that alter important amino-acids and, thus, have deleterious effects on protein structure or that affect expression (Nei 1987). Purifying selection can be indicative of the deleterious consequences of a given allele (Zhao e*t al*. 2003). Reduced genetic diversity at SNP loci and increased genetic distances between populations may be an indication of purifying selection (see below).

3.2.9.11.1 Genetic diversity

Estimation of gene diversity (heterozygosity), at SNP loci in the combined African resequencing samples (Figure 22), revealed differences with respect to gene diversity (although not significant) between nonsynonymous and synonymous SNP sites in the combined sample (Figure 22) as well as in the NIEHS sample (results not shown). Synonymous SNP sites in exons had the highest observed gene diversity, 0.40, followed by intronic SNP, which had a gene diversity of 0.31, whereas the mean gene diversities of nonsynonymous SNP sites were noticeably lower. Nonsynonymous SNP sites causing a conservative amino-acid change had a mean gene diversity of 0.19, whereas nonsynonymous SNP sites causing a radical-amino acid change had a mean gene diversity of 0.11. The lowest mean gene diversity (0.07) was for nonsense SNP sites, which result in a stop-codon. The difference in mean gene diversity between all SNPs compared with SNPs causing a radical-amino acid changes, was not significant (*P*>0.1).

A comparison of genetic diversity in 23238C allele SNP categories with that in 23238T alleles (Figure 23), revealed similar results to those observed for the combined African sample, in that differences with respect to gene diversity (although not significant) were observed between nonsynonymous and synonymous SNP sites in both the 23238C and the 23238T alleles. Intronic SNPs had a high observed gene diversity in both groups (0.36 in 23238C alleles and 0.33 in 23238T alleles), and nonsense SNPs had the lowest gene diversity (0.03 in 23238C alleles and 0.01 in 23238T alleles). 23238C alleles had a slightly lower frequency of nonsynonymous SNP sites causing radical-amino acid changes 0.19 compared to 0.22 in the 23238T alleles (Figure 23).

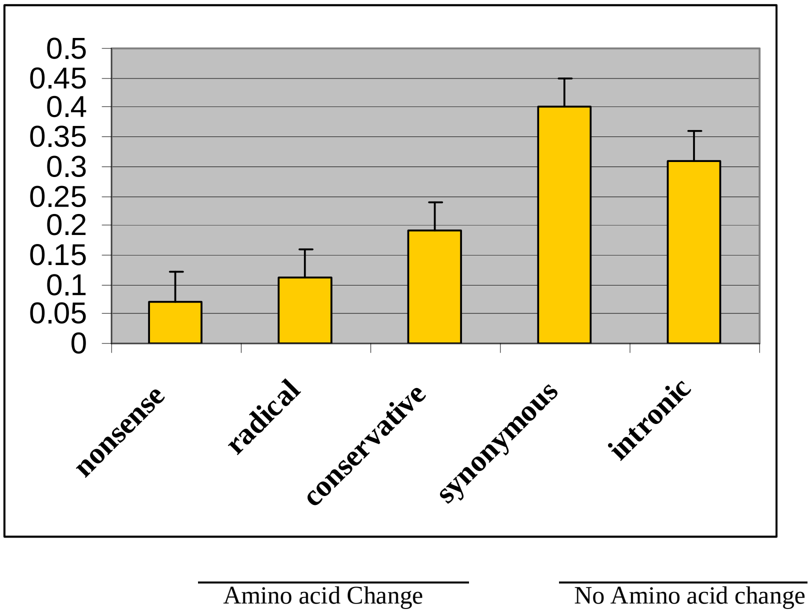


Figure 22. Mean gene diversity (heterozygosity) at SNP sites categorized by location in the gene and effect on protein function for the combined African resequencing samples. Error bars indicate standard errors. All means were not significantly different from that for nonsynonymous radical SNPs.

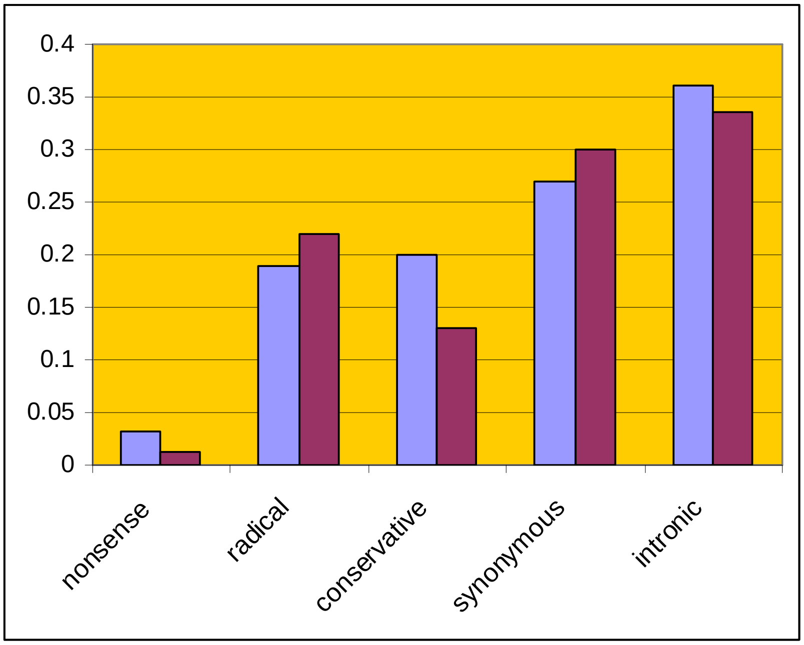


Figure 23. Mean gene diversity (heterozygosity) comparison between 23238C and 23238T alleles at SNP sites categorized by location in the gene and effect on protein function. All means were not significantly different from that for nonsynonymous radical SNPs. 23238C alleles are represented by blue bars, whereas the 23238T alleles are represented by purple bars.

3.2.9.11.2 Testing for natural selection at conservative nonsynonymous SNP sites

Sequences of *FMO2* of human and chimp (*Pan troglodytes*) and human and mouse (*Mus musculus*), taken from Ensembl, were aligned using the ClustalX, v.2.0.11. It was assumed that the chimp and mouse sequences represented the more common allele in these species. Following the maximum parsimony principle (Fitch 1971), the assumption that the amino-acid residue that represents the more common allele in all three species: human, chimp and mouse, has been conserved since the most recent common ancestor of these species.

Results showed that for the seven nonsynonymous SNPs [g.107A>G (D36G), g.7695T>A (F69Y), g.18237G>A (R238Q), g.19679A>G (E314G), g.19910G>C (R391T), g.22060T>G (N413K) and g.23115A>C (N431T)] causing a conservative amino-acid change, the non-derived allele was the same in all three species, which indicates that FMO2 amino-acid residues at these SNP positions are highly conserved among species.

3.2.9.11.3 Testing for natural selection at radical nonsynonymous and nonsense SNP sites

The same procedure was performed for radical and nonsense nonsynonymous SNPs (see above). Results showed that for the three nonsynonymous SNPs causing a radical amino acid-change [g.7731T>C (F81S), g.13693T>C (F182S) and g.13732C>T (S195L)], only one, g.13732C<T (S195L), had a common allele in humans that differed from the allele found in both chimp and mouse. Of the three SNPs resulting in a stop-codon (nonsense): g.10951delG (V113fsX), g.18269C>T (R249X) and g.23238C<T (Q472X), only the last had a common allele in humans different from the one in chimp and mouse (i.e., the allele common in humans is the derived allele (T), whereas the non-derived allele (C), is

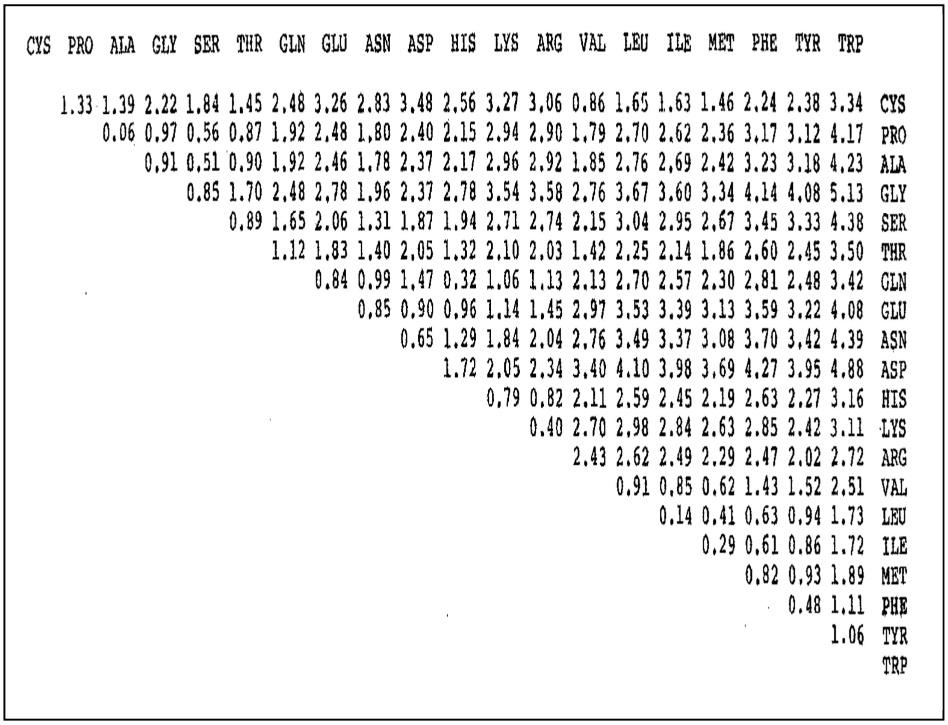
present in chimp and mouse. These results indicate that purifying selection is acting on the *FMO2* locus and that is why many amino-acid residues are conserved among species.

3.2.9.11.4 Genetic Distance

A genetic distance is a measure of the differences between populations. Genetic distance is measured using a variety of parameters. Genetic distance was used to compare the genetic distance between world populations, west- and east-African populations. A value of 0 indicates that two populations are genetically identical, whereas a value of 1 indicates that two populations are different species. Pairwise F_{ST} was used to estimate genetic distance between two populations (Reynolds *et al.* 1983), whereas hierarchical F_{ST} was used to estimate genetic distance within and between more than two populations (Excoffier *et al.* 1992).

SNPs were classified according to their location and effect on FMO2 function into: 1) SNPs in introns, 2) exonic synonymous SNPs, 3) exonic SNPs that cause conservative amino-acid changes, 4) exonic SNPs that cause radical amino-acid changes and 5) exonic nonsense SNPs causing premature stop-codons. The categorization of the nonsynonymous SNPs into conservative and radical was based on whether the SNP caused an amino-acid replacement involving two amino acids with a pairwise stereochemical difference >3, according to the scale of Miyata et al (1979) (Miyata, *et al.* 1979) (Figure 24)**.** Radical nonsynonymous changes included three SNPs (g.7731T>C in exon 2, causing F81S; g.13693T>C, causing F182S and g.13732C>T, causing S195L, both in exon 3). The F_{ST} values between pairs of SNPs were all close to zero (results not shown).

Figure24. Stereochemical differences between amino-acid pairs based on amino-acid residue and volume



CYS	PRO	ALA	GLY	SER	THR	GLN	GLU	ASN	ASP	HIS	LYS	ARG	VAL	LEU	ILE	MET	PHE	TYR	TRP	
	1.33	1.39	2.22	1.84	1.45	2.48	3.26	2.83	3.48	2.56	3.27	3.06	0.86	1.65	1.63	1.46	2.24	2.38	3.34	CYS
		0.06	0.97	0.56	0.87	1.92	2.48	1.80	2.40	2.15	2.94	2.90	1.79	2.70	2.62	2.36	3.17	3.12	4.17	PRO
			0.91	0.51	0.90	1.92	2.46	1.78	2.37	2.17	2.96	2.92	1.85	2.76	2.69	2.42	3.23	3.18	4.23	ALA
				0.85	1.70	2.48	2.78	1.96	2.37	2.78	3.54	3.58	2.76	3.67	3.60	3.34	4.14	4.08	5.13	GLY
					0.89	1.65	2.06	1.31	1.87	1.94	2.71	2.74	2.15	3.04	2.95	2.67	3.45	3.33	4.38	SER
						1.12	1.83	1.40	2.05	1.32	2.10	2.03	1.42	2.25	2.14	1.86	2.60	2.45	3.50	THR
							0.84	0.99	1.47	0.32	1.06	1.13	2.13	2.70	2.57	2.30	2.81	2.48	3.42	GLN
								0.85	0.90	0.96	1.14	1.45	2.97	3.53	3.39	3.13	3.59	3.22	4.08	GLU
									0.65	1.29	1.84	2.04	2.76	3.49	3.37	3.08	3.70	3.42	4.39	ASN
										1.72	2.05	2.34	3.40	4.10	3.98	3.69	4.27	3.95	4.88	ASP
											0.79	0.82	2.11	2.59	2.45	2.19	2.63	2.27	3.16	HIS
												0.40	2.70	2.98	2.84	2.63	2.85	2.42	3.11	LYS
													2.43	2.62	2.49	2.29	2.47	2.02	2.72	ARG
														0.91	0.85	0.62	1.43	1.52	2.51	VAL
															0.14	0.41	0.63	0.94	1.73	LEU
																0.29	0.61	0.86	1.72	ILE
																	0.82	0.93	1.89	MET
																		0.48	1.11	PHE
																			1.06	TYR
																				TRP

Miyata`s scale for pairwise stereochemical amino-acid differences. Figure taken from Miyata *et al.* (1979) (Miyata *et al.* 1979).

3.2.9.12 Testing for selection at intronic sequences

3.2.9.12.1 Pi (π) ratio comparisons for intronic sequences

Long-term persistence of both functional and non-functional alleles at the Flavin-containing monooxygenase 2 *(FMO2)* locus suggests that the locus may have been subjected to selective forces, i.e., the functional 23238C allele has been under purifying selection to remove deleterious SNPs. Balancing and positive selection are other selective forces that might be acting at the *FMO2* locus.

The positive selection hypothesis predicts that the level of polymorphism is lower in the selected allele than in the alternative SNP allele, in the proximity of the SNP [g.23238C>T (Q472X)] for example, because of the hitchhiking effect (Wang *et al.* 2006). π, nucleotide diversity per site, associated with the truncated and full-length allele, respectively, was used as an indication of the level of polymorphism (Wang *et al*. 2006).

One hypothesis is that the truncated 23238T alleles (null-alleles) may have been positively selected for and that the SNP that caused the truncation, g.23238C>T (Q472X), may be moving towards fixation in all human populations (pseudogenization). Positive selection of null alleles (i.e., pseudogenes), cannot be detected by a comparison of human and chimpanzee sequences. However, human population genetic data may retain signatures of selective sweeps for up to 200,000 years (Przeworski 2003). If pseudogenization is not yet complete, there is a good chance of being able to detect the signature of evolutionary forces responsible for the pseudogenization. If the nearly complete fixation of a null allele has been driven by positive selection there should be signals of recent (incomplete) selective sweeps, which could be detected by examining intra-specific variation of putatively neutral regions (e.g., intronic sequences) surrounding the null-allele (Wang *et al.* 2006). The LRH tests only detect recent selection (<30,000 years), which might be the reason Veeramah *et al*. (2008) did not find evidence of positive selection when testing HapMap data for *FMO2* (Veeramah *et al*. 2008). Analysis of resequencing data from regions close to the truncation SNP (within about 15 kb either side) may detect the remnants of a much older selective sweep.

The following was considered for the analysis:

Intronic segments situated within about 15 kb either side of the truncation SNP (about four upstream and four downstream) were used (see below). The number of SNPs in both the functional 23238C allele and the non-functional 23238T allele were compared by Fisher`s exact test. There was no significant difference between the frequency of SNPs on either a 23238C or a 23238T allele background in these intronic sequences, although several SNPs had a higher frequency on a 23238T allele background. π_T and π_C comparisons (by two-tailed Z test were also performed, results not shown). Sequence data available for *FMO2* from NIEHS was used for this analysis. Eight heterozygotes for the SNP g.23238C<T (Q472X) (C/Ts) (African-American and African) and 100 homozygote alleles (African, European, Hispanic and Asian) for the truncated 23238T allele (T/Ts) were used.

Length of intronic sequences used from (NIEHS) for π ratio comparisons were as follows:

1) Around the truncation SNP

+20478 to +21977 (located in the intron between exons 7 and 8), 1499 nucleotides.

+22078 to +23130 (located in the intron between exons 8 and 9), 1052 nucleotides.

+23479 to +24877 (downstream of exon 9), 1398 nucleotides.

2) Around exon 7

+18378 to +19527 (located in the intron between exon 6 and exon 7), 1149 nucleotides.

+19927 to +21027 (in the intron between exon 7 and exon 8), 1100 nucleotides.

3) Between exons 4, 5 and 6

+11128 to +13177 (in the intron between exon 4 and exon 5), 2049 nucleotides.

+13778 to +15827 (in the intron between exon 5 and exon 6), 2049 nucleotides.

4) Between exons 3 and 4.

+5528 to +7577 (in the intron between exon 3 and exon 4), 2049 nucleotides.

+7828 to +9327 (in the intron between exon 4 and exon 5), 1499 nucleotides

Tajima`s *D* and Fu and Li`s *D** and *F** test statistics for these intronic regions were calculated for the 23238C and 23238T allele groups. The 23238C allele group had a significantly positive Tajima`s *D* ($P$<0.05) for the region upstream of the g.23238C>T (Q472X) SNP, as well as significantly positive Fu and Li`s *D** and *F** test statistics (P<0.05). All these *P* values resulted from coalescent simulation, with intermediate recombination. These values may indicate population subdivision, since the 23238C alleles originated from African-American and African individuals. Whereas for the 23238T allele regions, in the sequence upstream of exon 4, there was a significantly positive Tajima`s *D* and Fu and Li`s *D** and *F** test statistics (P<0.05), (P<0.05) (from comparison with coalescent simulations that had the same sample size and recombination rate as the sample). The sequence downstream of exon 4 had a significantly positive Fu and Li`s *D*\* ($P$<0.05). Upstream of the truncation, there was negative Fu and Li`s *D** and *F** test statistics, although not significant. Downstream of the truncation SNP, Fu and Li`s *D** was significantly positive (P<0.001). Nucleotide diversity (π) was estimated for each set of sequences in African-American, European, Hispanic and Asian NIEHS individuals (Table 24).

The significantly positive values for Tajima`s *D* and Fu and Li`s *D** and *F** test statistics observed for any intronic sequences in the 23238T alleles may be due to population subdivision (F_{ST} = 0.20), since the 23238T alleles are from diverse populations (African-American, African, European, Hispanic and Asian populations). Comparisons between π_C and π_T were performed. Ratios of π_C/π_T were calculated across intronic sequences for African-American and African individuals from NIEHS (Figure 25). Since the only 23238C alleles came from heterozygotes in African-American and African individuals from NIEHS, only individuals with complete genotypes were used and haplotypes were inferred using Phase.

Table 24. π values for NIEHS populations across intronic sequences

Sequence range	**Population**			
	Af-Am and Af	**European**	**Asian**	**23238C**
5.527-7.577	**0.00346**	**0.00203**	**0.00223**	**0.00025**
7.828-9.327	**0.00243**	**0.00205**	**0.00235**	**0.00001**
11.128-13.177	**0.0015**	**0.00108**	**0.00108**	**0.00041**
13.778-15.827	**0.00248**	**0.00158**	**0.00245**	**0.00322**
18.378-19527	**0.00043**	**0.00036**	**0.00075**	**0.00001**
19.928-21.977	**0.00075**	**0.00075**	**0.00074**	**0.00026**
22.078-23.080*	**0.00184**	**0.0018**	**0.00197**	**0.00067**
23.479-24.877#	**0.00119**	**0.0011**	**0.00157**	**0.00001**

Table shows π values for NIEHS populations across eight intronic sequences. Af-Am, African American, Af, African, European and Asian 23238T alleles. *Intronic sequence upstream of g.23238C<T (Q472X) SNP, #intronic sequence downstream of g.23238C<T (Q472X).

Results indicate that π_C had lower values compared with π_T. This indicates lower nucleotide diversity in 23238C alleles compared with 23238T alleles (Table 24). Figure 25 displays the π_C/π_T ratios across intronic sites for African-American and African individuals. The lower values of nucleotide diversity in the 23238C alleles indicate the possibility than an incomplete selective sweep has been acting on the 23238C alleles with the resulting hitchhiking effect leading to a decrease in the nucleotide diversity of these alleles (see Figure 25).

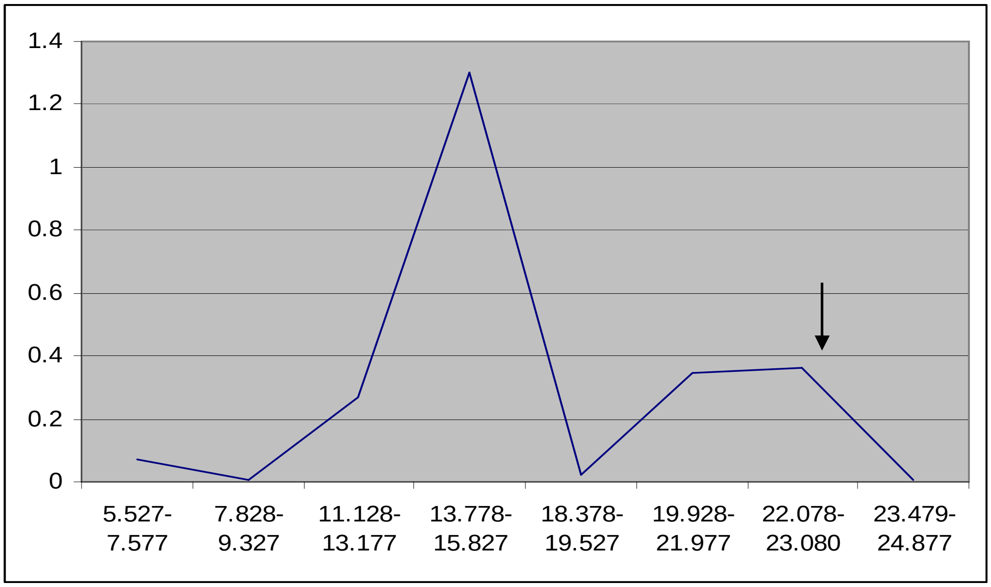


Figure 25. π_C/π_T for intronic sequences in African-American and African individuals. (Arrow indicates the location of the g.23238C>T (Q472X) SNP).

3.2.9.13 Testing for a common origin of African, European, Hispanic and Asian 23238T alleles

To determine whether there is a common origin of the African and the non-African T alleles, African-American, African NIEHS and combined west- and east-African homozygotes for the 23238T allele were combined and phased for 19 variants. Genotypes of European, Hispanic and Asian individuals from the NIEHS panel were phased for the same variants, in order to check for cosmopolitan haplotypes.

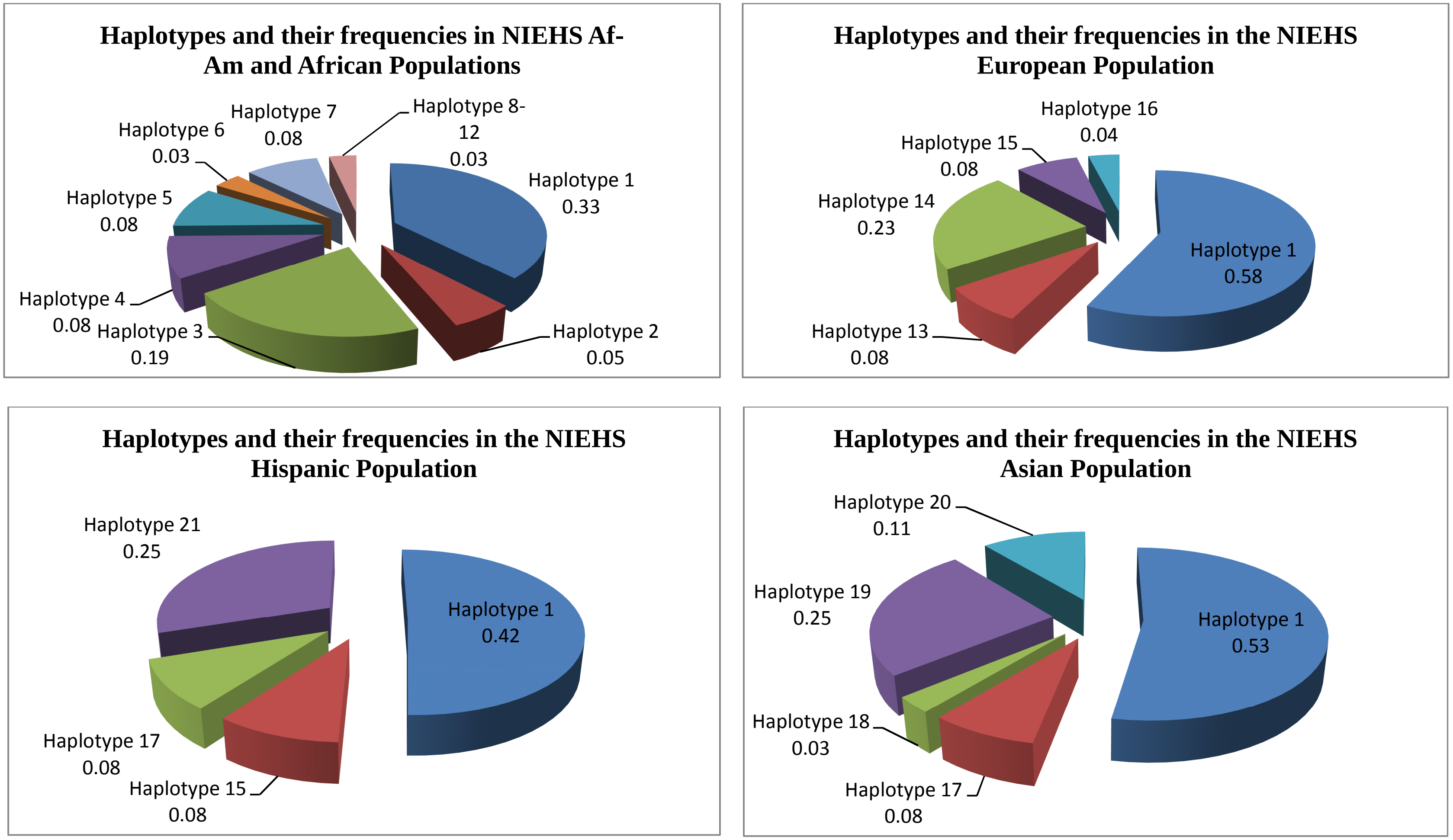
There was a common haplotype among all population groups, haplotype 1, which was present at a frequency ranging from 0.33 in Africans to 0.58 in Europeans. Africans shared only one haplotype (haplotype 1) with the other groups, whereas Europeans shared two haplotypes with Hispanics and, interestingly, Asian and Hispanic populations shared three haplotypes (Table 25 and Figure 26). These results indicate that 23238T alleles of the three old world populations may share a common ancestor.

Table 25. Haplotypes for the NIEHS populations

Haplotype id and occurence		+107A>G	+7696T>A	+7700-7702InsGAC	+7731T>C	+10951delG	+13693T>C	+13732C>T	+13733G>A	+18237G>A	+18269C>T	+19679A>G	+19839A>G	+19910G>C	+22026G>A	+22060T>G	+23087A>G	+23238C>T	+23300A>G	+23353-23354T	Frequency
Ancestral	0	A	T	del	T	G	T	C	G	G	C	A	A	G	G	T	A	C	A	del	0
Population Group																					
NIEHS African-Americans, Africans and Africans from the combined resequencing Samples (n=36)																					
1	12	A	T	del	T	G	T	C	A	G	C	A	A	G	G	T	A	T	A	del	0.33*
2	2	A	T	del	T	G	T	C	A	A	C	A	A	G	G	T	A	C	A	del	0.05
3	7	A	T	del	T	G	T	C	A	G	C	A	A	G	G	T	A	C	A	del	0.19*
4	3	A	T	del	T	G	T	C	A	G	C	A	A	C	G	T	A	C	A	del	0.08
5	3	A	T	del	T	G	T	C	A	G	C	A	A	C	G	T	A	T	A	del	0.08

6	1	A	T	del	T	G	T	C	A	A	C	A	A	G	G	T	A	T	A	del	0.03
7	3	A	T	del	T	G	T	C	A	G	C	A	G	G	G	T	A	T	A	del	0.08
8	1	A	T	del	T	G	T	C	A	G	T	A	G	G	G	G	A	T	A	del	0.03
9	1	A	T	del	T	G	C	C	G	G	C	A	G	G	G	T	A	T	G	del	0.03
10	1	A	T	del	T	G	C	C	G	G	C	A	G	G	G	T	A	T	A	del	0.03
11	1	A	T	GAC	T	G	T	C	A	G	C	A	A	G	A	T	A	C	A	del	0.03
12	1	A	T	GAC	T	G	T	C	A	G	T	A	G	G	G	G	A	T	A	del	0.03
NIEHS European Population (n=26)																					
1	15	A	T	del	T	G	T	C	A	G	C	A	A	G	G	T	A	T	A	del	0.58
13	2	A	T	del	T	G	T	C	A	G	C	A	A	G	G	T	A	T	G	del	0.08
14	6	A	T	del	T	G	T	T	G	G	C	A	A	G	G	T	A	T	A	del	0.23
15	2	A	T	del	T	G	C	C	G	G	C	A	A	G	G	T	A	T	G	del	0.08
16	1	A	T	GAC	T	G	T	C	A	G	C	A	A	G	G	T	A	T	A	del	0.04
NIEHS Hispanic Population (n=24)																					
1	10	A	T	del	T	G	T	C	A	G	C	A	A	G	G	T	A	T	A	del	0.42
15	2	A	T	del	T	G	C	C	G	G	C	A	A	G	G	T	A	T	G	del	0.08
17	2	A	T	del	T	G	T	T	G	G	C	A	A	G	G	T	A	T	G	del	0.08
21	6	A	T	del	T	G	T	T	G	G	C	G	A	G	G	T	A	T	A	T	0.25
NIEHS Asian Population (n=36)																					
1	19	A	T	del	T	G	T	C	A	G	C	A	A	G	G	T	A	T	A	del	0.53
17	3	A	T	del	T	G	T	T	G	G	C	A	A	G	G	T	A	T	G	del	0.08
18	1	G	T	del	T	G	T	T	G	G	C	A	A	G	G	T	A	T	A	del	0.03
19	9	A	T	del	T	G	T	T	G	G	C	G	A	G	G	T	A	T	A	T	0.25
20	4	G	T	del	T	G	T	T	G	G	C	A	A	G	G	T	A	T	A	del	0.11

Haplotypes and their estimated frequencies in the NIEHS Populations and the combined African sample. The nucleotide positions of SNPs are given relative to the A of the ATG translational initiation codon. Those that result in amino-acid substitutions, g.107 (D36G), g.7695 (F69Y), g.7700-7702 (71dupD), g.7731 (F81S), g.10951 (V113fsX), g.13693 (F182S), g.13732 (S195L), g.18237 (R238Q), g.18269 (R249X), g.19679 (E314G), g.19910 (T391R), g.22026 (E402K),g. 22060 (N413K), g.23238 (Q472X) and g.23353-23354 (F472fsX) are shown in bold. All polymorphisms are shown as derived changes in comparison with the human ancestral sequence. Shared haplotypes among populations are colour marked. *Indicates that the haplotype is functional, n=number of alleles in population. Nomenclature of haplotypes is the same as in Figure 26.



Haplotypes and their frequencies in NIEHS Af-Am and African Populations

Haplotypes and their frequencies in the NIEHS European Population

Haplotypes and their frequencies in the NIEHS Hispanic Population

Haplotypes and their frequencies in the NIEHS Asian Population

Figure 26. Haplotypes and their frequencies in the three old world populations.

3.2.10 Time to the most recent common ancestral sequence (T_{MRCA})

GeneTree was used to determine the time to the most recent common ancestral sequence (T_{MRCA}) as well as the ages of some of the SNPs observed in the combined African resequencing sample. Since GeneTree assumes an infinite-sites model of mutation without recombination, all recombinant haplotypes had to be excluded (Table 26). Therefore, of the ten distinct haplotypes present in the 30 African chromosomes (Table 16), only three could be used for this analysis, since an ancestral sequence needed to be included and this caused a reticulation in the network (diagram not shown), which indicates possible recombination. A median-joining network diagram was drawn for the non-recombinant haplotypes (Figure 27). GeneTree was used to estimate θ (a measure of the average expected per-site nucleotide diversity) using a maximum-likelihood (ML) method. θ is needed to calculate N_e (effective population size) since GeneTree gives times in units of $2N_e$ (see Sections 2.5.2.6 and 2.5.2.6.1). N_e is calculated from $\theta = 4\ N_{e,}\mu$. μ is the neutral mutation rate per region per generation and can be calculated from $\mu = \upsilon g L$, where υ is the neutral mutation rate per nucleotide per year, g is the generation time (20 years) and L is the number of silent sites (intronic + silent sites in exons), which was 3442 in the sample. $\theta_{\mathrm{ML}} = 2.8$, which is slightly lower than θ_{S}, the estimate from the number of segregating sites, which is 3.4. Another value required for the calculation of the neutral mutation rate is the net silent-site sequence divergence between humans and chimpanzee (i.e, divergence between humans and chimpanzee minus the divergence within these two species) and a human-chimpanzee divergence time of ~6 million years (Wall 2002).

The net silent-site sequence divergence was estimated using DnaSP and was = 33.3. $\upsilon = 8.3 \times 10^{-10}$/site/year. This value is slightly higher than estimates for many other human loci (Tishkoff and Verrelli 2003). So now N_e can be calculated, and was estimated to be 9,234 individuals, which is similar to the effective population size commonly found from studies of human genes (10,000 individuals) (Harding *et al*. 1997; Fullerton *et al*. 2002; Wooding *et al*. 2002; Tishkoff and Verrelli 2003). GeneTree gave estimates for T_{MRCA} and the ages of some of the variants present in the African sample in coalescent units of $2N_e$ generations. Each of the time estimates was converted into years using the estimate of N_e and a generation time of 20 years.

Figure 28 shows the GeneTree for African non-recombinant haplotypes. (Times in square brackets are the 95% confidence intervals for the times estimated using GeneTree, http://www.dimensionresearch.com). The time to the most-recent common ancestral sequence was estimated to be ~0.928 million years (myr) [0.843-1.012 myr]. T_{MRCA} (SD = 236 kyr). T_{MRCA} can also be calculated independently of GeneTree by using a coalescent model that assumes constant N_e. In such a model, the average pairwise divergence is dominated by the time taken for the last two lineages to coalesce, which is expected to take 2 N_e generations, or half the time to the most recent common ancestor (T_{MRCA}). Therefore, the average pairwise age of sequence diversity can be estimated from: human-chimpanzee divergence time x the average pairwise difference among human haplotypes/net sequence divergence between human and chimpanzee *FMO2* genes. This resulted in a T_{MRCA} = 1.3 million years (myr), which is slightly higher than the estimate from Genetree (0.928 million years).

GeneTree can also be used to estimate the coalescent times of haplotypes. The coalescent time for haplotypes 5, 6 and 9, which corresponds to the point at which these lineages coalesce to a common ancestor, is estimated as 732 kyr [662-

801 kyr] (SD = 193 kyr). Results indicate that 11 of the 12 variants predated the out-of-Africa migration. The SNP g.19839A>G (A367A) occurred ~0.83 million years ago (myr) [0.750-0.910 myr] (SD = 216 kyr). The only SNP estimated to have occurred after the migration of modern *Homo sapiens* out of Africa ~60,000 years ago is g.13732C>T (S195SL). The GeneTree time estimate for this SNP seems to be underestimated (45 kyr) [31.40-58.60 kyr] (SD = 38 kyr), because if it had arisen after modern *Homo sapiens* migrated out of Africa, it should not be present in Africa, but since the SNP is also observed in Africa at high frequency (0.53), which indicates that it has arisen in Africa before the migration. This is confirmed by the SNP being observed in all three old world populations. The SNP is present in the HapMap-CEU and the EGP-CEPH Panels at a frequency of 0.24 and 0.30, respectively, whereas it occurs at an average frequency of 0.52 in the combined HapMap-HCB and HapMap-JPT as well as in the EGP-Asian Panel. Five of the 12 SNPs [+7586C>T, +7883T>A, g.13732C>T (S195L), g.13733G>A (S195S) and +13952C>A] (Table 26) were also observed in Asian and European individuals, whereas another two [g.7731T>C (F81S) and g.18237G>A (R238Q)], were observed only in Europeans but not in Asians, which can give valuable information as to which haplotypes migrated out of Africa and which did not. Alternatively, haplotypes not observed in Asia or Europe might have migrated out of Africa, but did not survive in other environments.

Table 26. Non-recombinant haplotypes and their frequency

Haplotype id and occurrence	+7586C>T	**+7731T>C**	+7883T>A	**+10951delG**	**+13732C>T**	+13733G>A	+13952C>A	+17928C>T	**+18237G>A**	+18390T>C	+19839A>G	+19969G>A	
SNP id	1	2	3	**4**	**5**	6	7	8	**9**	10	11	12	**Frequency**
Ancestral (1) 0	C	**T**	T	**G**	**C**	G	C	C	**G**	T	A	G	**0.00**
5 12	.	**.**	.	**.**	**.**	A	A	T	**.**	C	.	A	**0.40**
6 2	.	**.**	.	**.**	**.**	A	A	T	**A**	C	.	A	**0.066**
9 3	T	**C**	A	**del**	**T**	.	.	.	**.**	.	G	.	**0.10**

Table displays the haplotypes used for GeneTree analysis. Their frequencies are displayed in the first column. There were 12 variants. Those resulting in an amino-acid change [g.7731T>C (F81S), g.10951delG (V113fsX), g.13732C>T (S195L) and g.18237G>A (R238Q) are shown in bold. The number under the SNP nucleotide position is the SNP id number as in the MJ network diagram (Figure 27) and in the GeneTree (Figure 28).

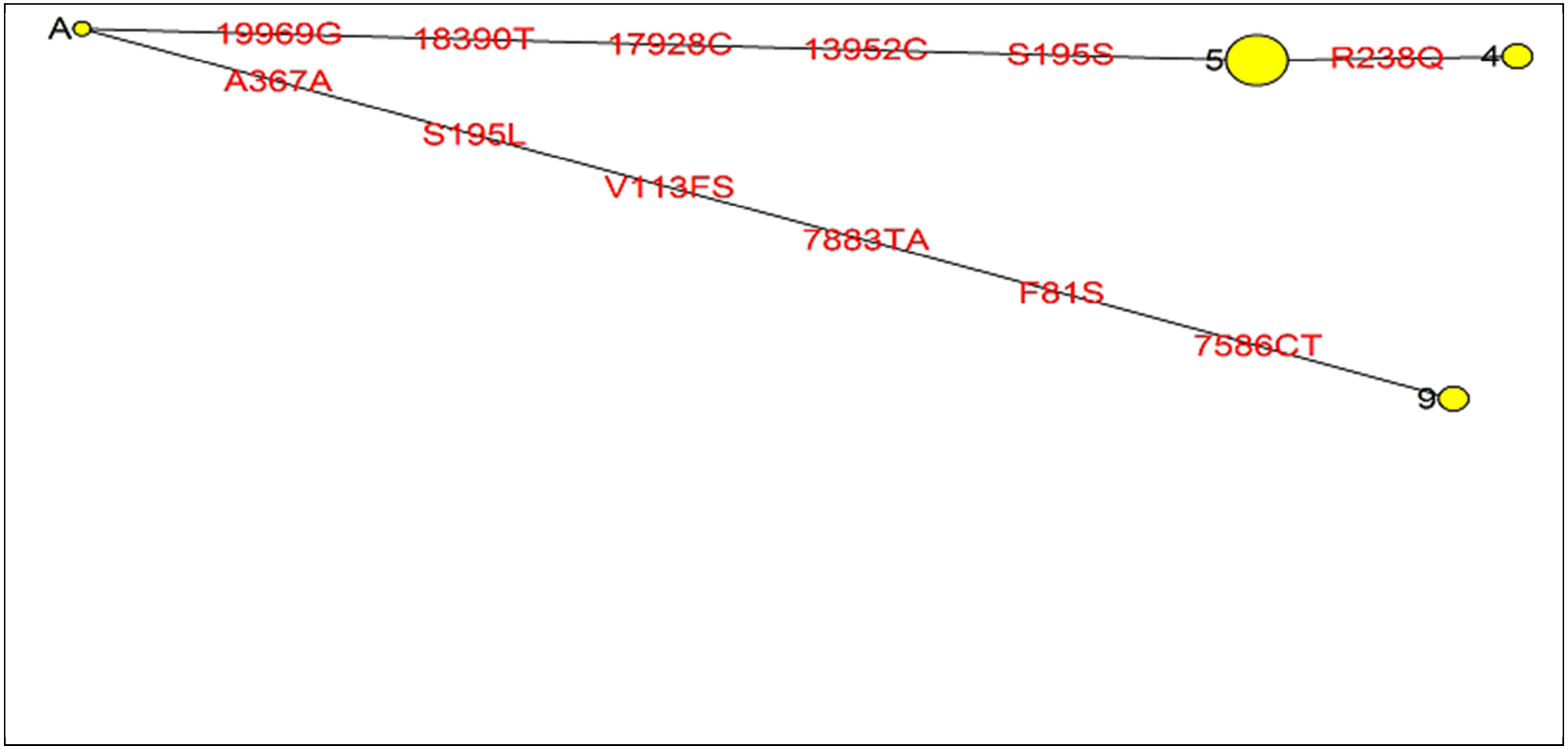


Figure 27. Median-joining (MJ) network diagram of the non-recombinant haplotypes used for GeneTree. Each of the three unique haplotypes is represented by a circle, the size of which is proportional to haplotype frequency. The lines connecting the haplotypes indicate the mutational relationship. Mutational differences between haplotypes are indicated on the branches of the network. Amino-acid residue positions on the protein identify mutations. Where multiple mutations occur along a branch, their order is arbitrary.

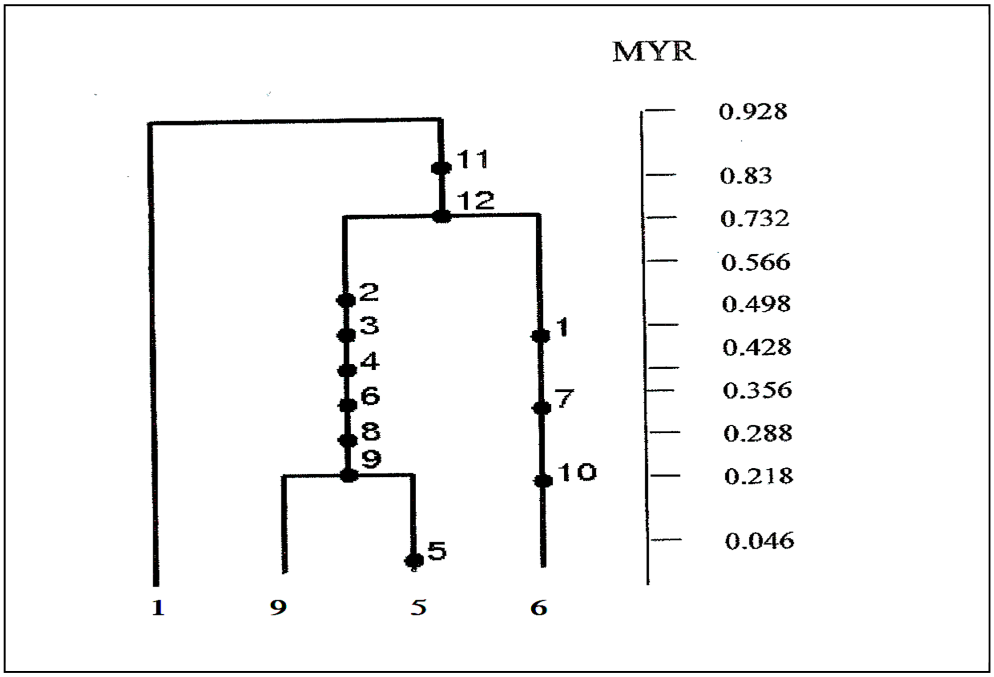


Figure 28. GeneTree diagram. The GeneTree diagram displays the times of some of the mutations and the coalescence times of haplotype clusters.

3.2.11 The evolutionary relationship among *FMO2* haplotypes

The evolutionary relationship among nine of the ten *FMO2* haplotypes for the combined African resequencing sample of 30 chromosomes, was visualized by a reduced-median (RM) network (Figure 29). Haplotype identities are the ones displayed in (Table 16), except for haplotype 10, which was removed from the diagram. The diagram displays each of the 9 haplotypes as a circle, the size of which is proportional to haplotype frequency.

These circles are connected by lines (branches). Mutations are indicated on these branches. When more than one mutation occurs on a branch, these mutations cannot be ordered and, therefore, their order is arbitrary along the branch. Since no haplotype corresponds to the ancestral sequence, the chimpanzee sequence is used as the root haplotype of the network. The diagram shows that the earliest mutation to occur is the g.23238C>T (Q472X) truncation SNP and that all other mutations occur on a 23238T allele background, except for g.18237G>A (R238Q), which can also occur on a 23238C allele background. The *FMO2* haplotypes fall into four main groups. The first group comprises of the ancestral haplotype, the second is made up of haplotypes 2 and 9. These haplotypes are connected by a node, and are separated by two c.SNPs: g.22060T>G (N413K) and g.23238C>T (Q472X).

The third group is separated from the second by 12 mutations and is made up of four haplotypes: haplotypes 1, 4, 5 and 6. Each of these haplotypes is separated by a single mutation: haplotypes 1 and 6, as well as haplotypes 4 and 5, are separated by the c.SNP g.23238C>T (Q472X), whereas haplotypes 1 and 4 and haplotypes 5 and 6 are separated by the c.SNP g.18237G>A (R238Q). The placement of haplotype 1 is ambiguous, as indicated by a reticulation in the network. Reticulations are due to homoplasy (the occurrence of evolutionary events leading to the formation of the same allelic state at a variable site more than once). This homoplasy is observed for g.18237G>A (R238Q) and g.23238C>T (Q472X). Homoplastic sites can be brought about either by recurrent mutation or by recombination. Haplotype 1 may have arisen via a recurrent mutation of haplotype 5 at position 18237G>A (R238Q). However, this site is not part of a CpG dinucleotide and therefore, will not be subjected to a higher mutation rate. A more plausible explanation would be that it arose via a single recombination event that took place between haplotype 5 and either haplotype 4 or 6. This is supported by the four-gamete test, which predicted a recombination event between sites g.18237G>A (R238Q) and g.23238C>T (Q472X). Four further mutations separate the third group from the fourth group. This group consists of three haplotypes, haplotypes 3, 7 and 8. Haplotypes 3 and 7 are separated by a single c.SNP g.23238C>t (Q472X), whereas haplotypes 7 and 8 are separated by an intronic SNP, +19969G>A.

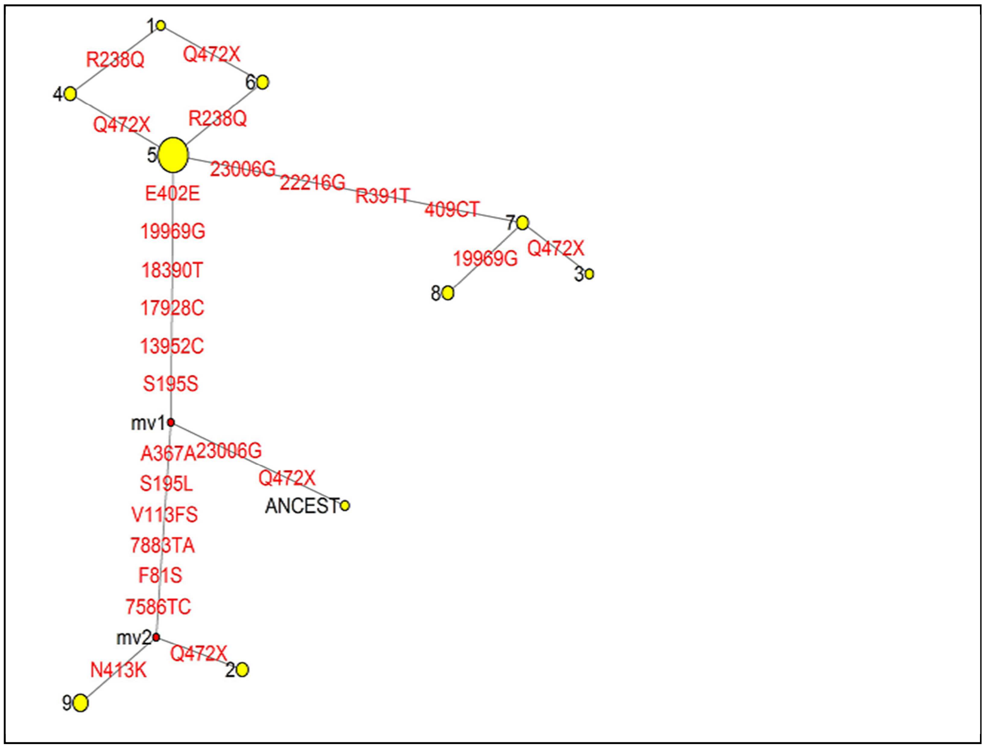


Figure 29. Reduced-median (RM) network of *FMO2* haplotypes for the combined African group of 30 chromosomes. Each of the nine unique haplotypes is represented by a circle, the size of which is proportional to haplotype frequency. The lines connecting the haplotypes indicate the mutational relationship. Mutational differences between haplotypes are indicated on the branches of the network. Amino-acid residue positions on the protein identify mutations. Where multiple mutations occur along a branch, their order is arbitrary. Haplotype 10 was removed.

4. General Discussion and Conclusion

4.1 Functional FMO2 is found at high frequency in east and west-Africa

The frequency of full-length 23238C alleles in east and west-Africa ranges from 0.18-0.21, with 0.32-0.36 of individuals possessing at least one full-length allele. Resequencing analysis of *FMO2* of 14 individuals homozygous for full-length 23238C alleles revealed that at least 0.54 of the full-length alleles encoded a functionally active protein and that 0.72 of these individuals possessed at least one potentially functional 23238C allele. These results indicate that the frequency of functional *FMO2* alleles and the proportion of individuals who possess at least one functional allele is relatively high in both east and west-Africa (~0.10 and ~0.18, respectively), which may have important implications for response to therapeutic drugs, especially thiourea-containing drugs used for the treatment of tuberculosis, which is widespread in Africa. The exact effect of FMO2 on these drugs is still unclear. The frequency of individuals in east and west-Africa (0.18 and 0.20, respectively) that possess potentially functional FMO2 was not significantly different (*P*>0.2).

Results show that the presence of full-length 23238C alleles in an individual, are not sufficient to indicate potentially functional FMO2, and that haplotypes need to be assessed for the absence of other nonsynonymous SNPs, and only then can potentially functional FMO2 be identified.

The results demonstrate that human *FMO2* is not a pseudogene, because functional FMO2 is still found at relatively high frequency. My results are similar to the results of the study by Veeramah *et al.* (2008) (Veeramah *et al.* 2008) in that the frequency of full-length 23238C alleles is relatively the same across sub-Saharan Africa but, interestingly, my results show that although there is no significant difference in the frequency of full-length 23238C alleles among nine population groups from sub-Saharan Africa, the frequency of potentially functional 23238C alleles differed significantly among these populations (P<0.05).

Another advantage of the study performed in this thesis is that in the study by Veeramah *et al*. (2008) samples were only genotyped for the g.23238C>T (Q472X) SNP and not for any other nonsynonymous SNPs, which were reported by other studies to cause loss of function (e.g., g.10951delG (V113fsX) and g.13732C>T (S195L) (Furnes *et al* 2003; Krueger *et al*. 2005, see Table 1), whereas this thesis considered testing the full-length 23238C alleles for the presence of nonsynonymous SNPs.

Although a recent study of single-strand conformation polymorphism (SSCP) of 50 African-Americans reported a number of previously unknown *FMO2* variants, no attempt was made to establish the background on which the SNPs occurred (Furnes *et al*. 2003). To address this point, I initially genotyped individuals from nine sub-Saharan African population groups for five SNPs, which were reported, by NCBI, to be at high-frequency in Africa. Genotyping results indicated that the frequency of full-length 23238C alleles that are potentially functional ranged from 0.35 to 1.00. The frequency of potentially functional *FMO2* alleles, in comparison to the total number of alleles, in the nine African population groups ranged from as low as 0.08 in Nuer to 0.35 in Mambila.

In an attempt to confirm these promising results, samples from west- and east-Africa were first genotyped for the g.23238C and g.23238T alleles and then resequenced to test for the occurrence of any other deleterious SNPs that might occur on a full-length 23238C allele background. Resequencing results confirmed that potentially functional FMO2 is at high frequency in sub-Saharan Africa, since more than half (0.55) of the full-length 23238C alleles are not associated with a nonsynonymous SNP (see Table 17). Resequencing revealed no novel SNPs. According to the equation $(1-p)^n$ = significance threshold, where n is the number of alleles and p is the proportion of the derived allele of a given variant, the minimum number of alleles needed to identify the proportion of derived alleles with a 95% confidence when the derived allele frequency is 0.03, 0.04 and 0.05 is 98, 73 and 58 respectively. So the minimum number of individuals needed for the identification of derived alleles with 95% confidence at frequencies of 0.03, 0.04 and 0.05 is 49, 36 and 29, respectively. So the African resequencing samples which consist of 96 alleles are large enough to identify derived allele frequencies with 95% confidence at frequencies of 0.04 or above.

4.1.1 Comparison of previously reported *FMO2* haplotypes with the ones observed in this thesis

The study by Krueger *et al.* (2005) compared the derived allele frequency of SNPs in Hispanic-Americans (28 23238C/T heterozygotes and one 23238C homozygote) with the ones observed for the African-American (50 individuals) study by Furnes *et al.* (2003) (Furnes *et al.* 2003). Krueger *et al.* (2005) genotyped the Hispanic population for four SNPs, which were the most frequent in the population of 50 African-Americans: the duplication at g.7700-7702dupGAC (71Ddup), g.10951delG (V113fsX), g.13732C>T (S195L) and g.22060T>G (N413K) and haplotypes were inferred. Twenty-one of these Hispanic individuals (0.72) possessed a full-length *FMO2* allele that had no other SNPs associated with it, and, therefore, are predicted to express functional FMO2. In contrast, the remaining eight (0.28) were heterozygotes for g.13732C>T (S195L). About 0.10 of the latter were also heterozygous for two more SNPs: g.10951delG (V113fsX) and g.22060T>G (N413K).

In order to compare the results from the Krueger *et al.* study with the study performed in this thesis, 23238C and 23238T homozygotes were phased for the four SNPs mentioned above. Of the 24 23238C and 24 23238T homozygotes, haplotypes of 17 of each type of homozygote were successfully inferred using Phase. The proportion of homozygous 23238C individuals that did not possess any of the four SNPs mentioned above was (0.71), 0.12 had g.13732C>T (S195L) present, whereas 0.06 had two SNPs present: g.13732C>T (S195L) and g.22060T>G (N413K). Krueger *et al*. (2005), also genotyped 124 Puerto Rican-American 23238T homozygote individuals. Of these, 0.31 had none of the other SNPs associated, whereas 0.67 had the g.13732C>T (S195L) SNP associated with the 23238T allele. Twenty-seven percent were also homozygous for the g.13732C>T (S195L) SNP.

In the study by Krueger *et al.* (2005), the g.10951delG (V113fsX) SNP was found only in individuals who also had g.13732C>T (S195L) present. The same individuals also expressed the g.22060T>G (N413K) SNP. The latter SNP was also present alone in 0.02 of the individuals. None of the Hispanic individuals genotyped possessed the g.7700-7702dupGAC (71Ddup) SNP. In comparison, for the 23238C and T homozygote individuals included in this thesis, 0.82 of the 23238T homozygotes had none of the other four SNPs present and 0.12 possessed the g.13732C>T (S195L) SNP. One similarity to the Krueger *et al*. (2005) study is the absence of the duplication SNP from both the 23238C and 23238T alleles.

None of the 23238C or the 23238T homozygotes possessed the g.10951delG (V113fsX) SNP.

The differences in the frequency of non-functional 23238T alleles observed in the nine African population groups (see Table 9) may be indicative of either an old selective sweep or it may be explained by mutations occurring across different regions of the 23238T allele.

4.2 Homozygote increase in genotyping samples

Significant deviation from HWE was observed for three of the SNPs [g.107A>G (D36G), g.10951delG (V113fsX) and g.23238C>T (Q472X)] in the genotyping samples. In Gurage, for g.107A>G (D36G), (*P*<0.005) and for g.10951delG (V113fsX), (*P*<0.001), in Fulbe, for g. 107 A>G (D36G), (*P*<0.001), in Manjak, for g.10951delG (V113fsX) (*P*<0.005) and g.23238C>T (Q472X) (*P*<0.05), and in Nuer, for g.23238C>T (Q472X) (*P*<0.005). The deviation was caused by less than the expected number of heterozygotes. This might be caused by the structure of these populations. Inbreeding may be causing the decrease in heterozygotes, since inbreeding leads to an observed increase in homozygotes (Charlesworth and Charlesworth 1987).

4.3 The evolution of *FMO2*

4.2.1 The mutational relationship of *FMO2* variants

4.2.1.1 For the genotyped populations

The reduced-median network diagram (see figure 13) showed that mutations occur on both a full-length 23238C and a truncated 23238T allele background. More complicated network diagrams for some populations were observed, due to homoplasy, which may be explained by recombination, since the occurrence of mutations on a full-length 23238C allele background is very low.

4.2.1.2 For the combined African resequencing sample

The reduced-median network diagram (Figure 29) for the combined African sample supports the results for the genotyped samples from the nine African population groups in that it shows that mutations mainly occur on a truncated 23238T allele background, except for the SNP g.18237G>A (R238Q), which occurs on both a truncated 23238T and a full-length 23238C allele background. The occurrence of this SNP on a full-length 23238C allele background can be explained by recombination. This is supported by the results of the four-gamete test, which predict a recombination event between g.18237G>A (R238Q) and g.23238C>T (Q472X). The diagrams (both for the genotyping populations and for the combined African sample) indicate that all of the other SNPs occurred after the g.23238C>T (Q472X) SNP, which was estimated by maximum-likelihood coalescent analysis to have occurred some 500 thousand years ago (kyr) (Veeramah *et al*. 2008).

4.2.2 Recombination and linkage disequilibrium

The presence of sites with four-gametes indicates either recombination or recurrent mutation. Since the possibility of recurrent mutation at the same site is very small, recombination is the more plausible explanation. The four-gamete test results predict five recombination events for the combined African resequencing sample (see Section 3.2.7). Despite the number of recombination events, strong significant LD is observed for both the 23238C and 23238T alleles. Significant LD for the 23238T allele group SNPs extends for 23kb, whereas the LD for the 23238C allele group SNPs extends for 22.6kb along the FMO2 gene. Significant

LD was also observed for the NIEHS Asian (22.2kb) and European (22.1kb) groups. The significant LD for the NIEHS Asian group was supported by values of Z_{nS} and Z_a that were both lower and higher than expected under a neutral model of evolution ($P<0.05$). The extensive LD observed in the NIEHS Asian group is consistent with a recent bottleneck (Becquet 2003), since recent, long bottlenecks favour the complete coalescence of most lineages during the bottleneck.

4.4 Natural selection

Enzymes that metabolize foreign chemicals are at the interface between an organism and its chemical environment and, thus, are likely candidates for natural selection (Verrelli *et al*. 2002)

4.4.1 Examining *FMO2* for evidence of natural selection

The LRH test used by Veeramah *et al.* (2008) (Veeramah *et al.* 2008) on human SNP data from HapMap phases 1 and 2 did not detect positive selection for either the 23238C or 23238T *FMO2* alleles. However, the LHR test can only detect recent positive selection (20,000-30,000 years) and, consequently, would have missed a more ancient selection event. Several population genetic methods that test DNA sequence variation for signatures of selection indicate that the *FMO2* locus might have been subjected to selection. Extensive LD is observed for both 23238C and T allele groups as well as the NIEHS Asian and European groups. The allele-frequency spectrum for all these groups shows an excess of intermediate-frequency variants, when compared with a neutral model (Figures 20a, b, c and d). Significantly positive Tajima`s *D* for the NIEHS Asian group supports the presence of an excess of intermediate-frequency, compared with low-frequency, variants.

Measures of population structure on a global level indicate that only ~0.10-

0.16 (Wright’s fixation index F_{ST} = (0.10-0.16) of observed genetic variation is due to differences among populations from Africa, Europe and Asia (Tishkoff and Verrelli 2003; International HapMap Consortium 2005; Weir *et al*. 2005). My result is higher than this range, F_{ST} = 0.20. The NIEHS Asian group shows an increased number of derived alleles present at frequencies reaching 0.27. The exposure to new chemical environments out of Africa may have rendered expressing the 23238C allele less favourable. This may be the reason that 23238C alleles are not found in Asia, although this may be because modern *Homo sapiens* who migrated into Asia did not possess 23238C alleles, or that these alleles were eliminated during the bottleneck that occurred after the migration out of Africa.

4.4.1.1 The McDonald-Kreitman test

The results of the MK test for the combined African sample as well as for both the 23238C and the 23238T allele groups showed an excess of polymorphic nonsynonymous SNPs, although the results were only significant for the combined (*P*<0.03) and the 23238T allele groups (*P*<0.03) and not for the 23238C allele group (*P*>0.05).

4.4.1.2 Purifying selection and genetic distance between SNP categories

It is expected that purifying selection acting to eliminate deleterious mutations is accompanied by directional selection on a favorable allele, which, in turn, results in hitchhiking events on linked neutral polymorphisms.

Purifying selection eliminates deleterious mutations, resulting in lower levels of genetic variation. Using Tajima`s *D* test statistic and Fay and Wu`s *H* statistic in combination is useful for distinguishing between selective forces and demographic events. If, for example, a significantly negative value of the *D* statistic is accompanied by a significant *H* value, recent population expansion and background selection can be excluded (Kim and Stephan 2000). Although such a result is not observed for any of the groups analysed, reduced genetic diversity at *FMO2* SNP loci and increased genetic distances between populations can alternatively, indicate purifying selection.

4.4.1.3 Testing for selection at intronic sequences

Putatively neutral regions can be used to detect a selective sweep. It is evident due to the presence of potentially functional FMO2, that the pseudogenization of *FMO2* is not yet complete; therefore intronic sequences were examined to test for such a sweep. Several tests (Tajima`s *D* and Fu and Li`s *D** and *F** test statistics) as well as nucleotide diversity comparisons between 23238C and 23238T alleles were applied to intronic sequences in an attempt to detect selection. The 23238C alleles had noticeably less nucleotide diversity, in comparison with 23238T alleles. These neutrality tests had significantly positive values for various sequences (see Section 3.2.9.12.1). These results indicate that the 23238C allele might have been the subject of a selective sweep which resulted in a decrease in nucleotide diversity. The high Ka/Ks ratio (>1) and low nucleotide diversity for intronic 23238C allele sites, extensive LD, as well as the allele-frequency spectrum with an excess of intermediate-frequency variants may be indicate of a selective sweep acting on the 23238C allele.

The results observed for the NIEHS Asian group; extensive LD with a significantly lower than expected value for Z_{nS} and a significantly higher than expected Z_a and of various significant neutrality test values (positive Tajima`s *D*, positive Fu and Lis` *D* and *F* test statistics, negative Fay and Wu`s *H* test, as well as an allele-frequency spectrum displaying an excess of intermediate-frequency variants) indicate that Asians underwent a bottleneck after the out-of-Africa migration, and that the resulting decrease in population size, caused an increase in genetic drift which in turn, resulted in an increase in intermediate-frequency variants (Rosenberg *et al.* 2002).

4.5 A common origin for the 23238T alleles

My results provide information about the evolutionary relationships of mutations and haplotypes of *FMO2,* and a comparison between the non-functional 23238T alleles in the three old world populations (Africans, Europeans and Asians).

It is important to determine whether there is a common origin for the 23238T allele in the three old world populations. The presence of a common haplotype shared by all three population groups indicate that the 23238T alleles for all populations were similar, and that Asian and European 23238T alleles share a common ancestor with African 23238T alleles. These findings further support the conclusion that the g.23238C>T (Q472X) mutation predated the out-of-Africa migration of *Homo sapiens,* since all out-of-Africa 23238T alleles share haplotypes with the more diverse African 23238T alleles.

4.6 Time to the most recent common ancestral sequence (T_{MRCA})

4.6.1 Time depth of *FMO2* variation and ages of some mutations

Both GeneTree and an independent coalescent model were used for the estimation of T_{MRCA}. GeneTree gave an estimate of 0.928 million years [0.843-1.012 myr] (SD = 236 kyr), whereas the coalescent model resulted in a slightly higher estimate of 1.3 million years (myr). GeneTree was also used to date some of the *FMO2* mutations. GeneTree results indicate that 11 out of the 12 variants analysed by GeneTree predated the out-of-Africa migration, except for g.13732C>T (S195L). The frequency of this SNP in Africa is high, as reported by

NCBI in the HapMap-Yoruba Panel (0.53) and the EGP-Yoruba panel (0.58) as well as in the combined African resequencing sample (0.11), indicating that the age of the mutation estimated by GeneTree (45 kyr) [31.40-58.60 kyr] (SD = 38 kyr) might be low and that the SNP actually predated the migration of modern humans out-of-Africa. The presence of the SNP in all three old world populations supports this.

The information about which SNPs are found both in and out of Africa will be used in an attempt to determine which haplotypes left Africa and which did not. Of the 12 variants mentioned in Table 26, five were also observed outside Africa, in Asian and European individuals [+7586C>T, +7883T>A, g.13732C>T (S195L), g.13733G>A (S195S) and +13952C>A], whereas another two [g.7731T>C (F81S) and g.18237G>A (R238Q)] were observed only in Europeans but not in Asians. These results suggest that all three haplotypes displayed in Table 26 have migrated out of Africa, but must have recombined [e.g., haplotype 9 contains the [g.7731T>C (F81S) and the g.10951delG (V113fsX) SNP]. The former is observed only in Europeans, whereas the latter is not observed in Europeans and Asians, it is only observed in African-Americans at a frequency of 0.125

4.7 Concluding Statement

It is well documented that enzymes involved in the metabolism of foreign compounds exhibit genetic variation (Board *et al*. 1998; Bertilson *et al*. 2002)

The main findings of this Thesis are:

1) The frequency of full-length functional FMO2 is high in sub-Saharan Africa (at least 0.54), which has important implications for therapeutic

treatment of sub-Saharan Africans and individuals of recent African descent with drugs that are substrates of FMO2.

2) The frequency of full-length alleles is not significantly different in sub-Saharan Africa, whereas the frequency of inferred potentially functional alleles is significantly different among populations in sub-Saharan Africa (P<0.05).
3) The lower intronic diversity in g.23238C allele intronic sequences and a $K_a/K_s > 1$, indicate a possible selective sweep acting on the g.23239C alleles.

From the results observed in this thesis, it is possible to gain insights into the evolutionary history of the *FMO2* gene. The presence of full-length functional *FMO2* alleles at high frequency in Africa (≥54%) and in contrast, the complete absence of these functional alleles in European and Asians indicates that there might have been selection for the truncated allele in the latter populations and/or for the full-length allele in Africa (see finding 3). Alternatively, individuals possessing a full-length 23238C allele may not have been part of the migration out of Africa or the allele may not have survived the bottleneck associated with the migration. It is apparent that the human *FMO2* gene has been under the influence of complex forces leaving a distinct pattern of diversity.

5. Future Work

Sub-Saharan Africa has an important and rich history from which we can gain insights into the origin of our species and how genetic variation affects human phenotypes, including complex diseases and the response to drugs.

Future work should include sequencing more intronic sequences of *FMO2,* since positive selection cannot be detected by the comparison of human and chimp sequences and if the pseudogenization is not yet complete, there is a good chance of being able to detect the remnants of a selective sweep (if any), by examining intra-specific variation at putatively neutral regions (e.g., intronic sequences) surrounding the null-allele. It would also be useful to sequence a number of g.23238C>T (Q472X) homozygotes in an attempt to further investigate the background on which other nonsynonymous SNPs occur. More African individuals (n=150) should be sequenced for the complete coding region of *FMO2* in order to make sure that even very rare mutations (1%) are picked up at a confidence of 95%, in order to completely exclude any other nonsynonymous SNPs occurring on a g.23238C background either known or novel. Analysis of more independent loci and a larger number of African populations, particularly from east Africa, will be necessary to better estimate the number and source of migration events out of Africa (Reed and Tishkoff 2006).

Research is needed to examine the effect of various nonsynonymous SNPs on protein function. Another interesting field of research would be to investigate the potential correlation between *FMO2* genotypes and the metabolism of two commonly used second-line antitubercular drugs TAZ and ETA *in vivo*, as well as studying the various responses to these drugs. Knockout mice models are the best way to establish genotype to phenotype relations.

6. References

Adler, P and Pouwel, R.L (2007). World Civilizations: Since 1500. Volume 2. 5th Edition. Thomson Wadsworth. Boston.

Akerman, B.R, Forrest, S, Chow, L, Youil, R, Knight, M and Treacy, E.P (1999a). "Two Novel Mutations of the *FMO3* Gene in a Proband with Trimethylaminuria". Hum Mutat 13(5): 376-379.

Akerman, B.R, Lemass, H, Chow, L.M, Lambert, D.M, Greenberg, C, Bibeau, C, Mamer, O.A and Treacy, E.P (1999b). "Trimethylaminuria is Caused by Mutations of the *FMO3* Gene in a North American Cohort". Mol Genet Metab 68(1): 24-31.

Akey, J.M, Eberle, M.E, Rieder, M.J, Carlson, C.S, Schriver, M.D, Nickerson, D.A and Kruglyak, L (2004). "Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes". PLoS Biology 2(10): 1591-1599.

Alahari, A, Trivell, X, Guèrardel, Y, Dover, L.G, Besra, G.S, Sachettini, J.C, Reynolds, R.C, Coxon, G.D and Kremer, L (2007). "Thiacetazone, an Antitubercular Drug that Inhibits Cyclopropanation of Cell Wall Mycolic Acids in *Mycobacteria".* PLoS ONE 2(12): e1343.

Alfieri, A, Malito, E, Orru, R, Fraaije, M.W and Mattevi, A (2008). "Revealing the Moonlighting Role of NADP in the Structure of a Flavin-containing Monooxygenase". PNAS 105(18): 6572-6577.

Allerston, C.K, Shimizu, M, Fujieda, M, Shephard, E.A, Yamazaki, H and Phillips, I.R (2007). "Molecular Evolution and Balancing Selection in the Flavin-containing Monooxygenase 3 Gene *(FMO3)*". Pharmacogenet Genom 17(10): 827-839.

Atkinson, Q.D, Gray, R.D and Drummund, A.J (2008). "mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory". Mol Biol Evol 25(2): 468-474.

Ayesh, R, Mitchel, S.C, Zhang, A and Smith, R.L (1993). "The Fish Odour Syndrome: Biochemical, Familial and Clinical Aspects". Br J Med. 307:655-657.

Bamshad, M and S. P. Wooding (2003). "Signatures of Natural Selection in the Human Genome". Nat Rev Genet 4(2): 99-111.

Bandelt, H.J, Forster, P, Sykes, B.C and Richards, M.B (1995). "Mitochondrial Portraits of Human Populations Using Median Networks". Genet 141(2): 743-753.

Bandelt, H.J, Forster, P and Roehl, A (1999). "Median-joining Networks for Inferring Intraspecific Phylogenies". Mol Biol Evol 16(1): 37-48.

Barber, M, Conrad, M.E, Umbreit, J.N, Barton J.C and Moore, E.G (2000). "Abnormalities of Flavin Monooxygenases as an Etiology for Sideroblastic Anaemia". Am J Hematol 65: 149-153.

Barreteau, D, Breton, R and Dieu, M (1984). "Les langues. In le Nord du Cameroon: Des Hommes, Une Region. The Languages of North Cameroon: One Region". Collection Memoires 102: 159-80.

Bates, D (1979). The Abyssinian Difficulty: the Emperor Theodorus and the Magdala Campaign. Oxford University Press. Oxford.

Belldina, E.B, Huang, M.Y, Schneider, J.A, Brundage, R.C and Tracy, T.S (2003). "Steady-State Phamacokinetics and Pharmacodynamics of Cysteamine Bitartrate in Paediatric Nephropathic Cystinosis Patients". Br J Clin Pharmacol 56(5): 520-525.

Benedetti, M.S and Keith, F (2007). "Amino Oxidases and Monooxygenases in the *Vivo* Metabolism of Xenobiotic Amines in Humans: Has the Involvement of Amine Oxidases Been Neglected?" Fund Clin Pharmacol 21(5): 467-480.

Berman, S.B and Greenamyre, J.T 2006. " Update on Huntington`s Disease". Curr Neur Neuroscein Rep 6:281-286.

Bertilson, L, Dhal, M.L, Dalen, P and Al-Shurbaji, A (2002). "Molecular Genetics of *CYP2D2*: Clinical Relevance with Focus on Psychotropic Drugs". Br J Clin Pharmacol 53(2): 111-122.

Becquet, C (2003). Signatures of a Population Bottleneck can be Localised along a Recombining Chromosome. Report of 5th-year Internship. University of Edinburgh, Edinburgh.

Biaglow, J. E, Issels, R.W, Gerweck, L.E, Varnes, M.E, Jacobsen, B, Mitchell JB and Russo, A (1984). "Factors Influencing the Oxidation of Cysteamine and Other Thiols: Implications for Hyperthermic Sensitization and Radiation Protection". Radiat Res 100(2): 298-301.

Bielawski, J.P and Yang, Z (2000). "Statistical Methods for Detecting Molecular Adaptations". Trends Ecol Evol 15(12): 496-503.

Board, P, Blackburn, A, Jermiin, L.S and Chelvanayagam, G (1998). "Polymorphism of Phase 2 Enzymes: Identification of new Enzymes and Polymorphic Variants by Database Analysis". Toxicol Lett 102-103: 149-154.

Borbas, T (2006). "Insulin as a Regulator of Flavin-containing Monooxygenase Enzyme in Streptozotocin-induced Diabetic Rats". Semmelweis University. Budapest. PhD Thesis.

Brown, M.R, Fisher, L.A, Sawchenko, P.E, Swanson, L.W and Vale W.W 1983. "Biological Effects of Cysteamine: Relationship to Somatostatin Depletion". Regulat Pep 5(2): 163-179.

Brown, P (1992). "Cheap TB Drug 'Too Dangerous' for Africa". New Scientist 135(1836): 5.

Brunelle A.B, Y.A, Lin, J, Russel, B, Luy, L, Berkman, C and Cashman, J.R (1997). "Characterization of Two Human Flavin-containing Monooxygenase (Form 3) Enzymes Expressed in *Escherichia Coli* as Maltose Binding Protein Fusions". Drug Metab Dispos 25: 1001-1007.

Brunet, M.G, Pilbeam, D, Mackaye, HT, Likius, A, Ahounta, D, Beauvilain, A, Blondel, C, Bocherens, C, Boisserie, J.R, De Bonis, *et al*. (2002). "A new Hominid from the Upper Miocene of Chad, Central Africa". Nature 11(418):145-151.

Campbell, M.C and Tishkoff, S.A (2008). "African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins and Complex Disease". Ann Rev Genom Hum Genet 9: 403-433.

Carlson, C. S, D, Thomas, J, Eberle, M.A, Swanson, J.E, Livingston, R.J, Rieder, M.J and Nickerson, D.A (2005). "Genomic Regions Exhibiting Positive Selection Identified from Dense Genotype Data". Genom Res 15(11): 1553-1565.

Cashman, J.R (1995). "Structural and Catalytic Properties of the Mammalian Flavin-containing Monooxygenase". Chem Res Toxicol 8(2): 165-181.

Cashman, J.R (2000). "Human Flavin-containing Monooxygenase Substrate Specificity and Role in Drug Metabolism". Curr Drug Metab 1(2): 181-191.

Cashman, J.R, Y. A. Bi, Lin, J, Youil, R Knight, M, Forrest, S and Treacy, E (1997). "Human Flavin-containing Monooxygenase Form 3: cDNA Expression of the Enzymes Containing Amino Acid Substitutions Observed in Individuals with Trimethylaminuria". Chem Res Toxicol 10(8): 837-841.

Cashman, J.R, Lattard, V and Lin, J (2004). "Effect of Total Parenteral Nutrition and Choline on Hepatic Flavin-containing and Cytochrome P-450 Monoxygenase Activity in Rats." Drug Metab Dispos 32(2): 222-229.

Cashman, J.R, Xiong, Y.N, Xu, L and Janowsky, A (1999). "*N*-Oxygenation of Amphetamine and Methamphetamine by the Human Flavin-containing

Monooxygenase (form 3): Role in Bioactivation and Detoxification". J Pharmacol Exp Thr 288(3): 1251-1260.

Cashman, J.R and Zhang, J (2002). "Interindividual Differences of Human Flavin-containing Monooxygenase 3: Genetic Polymorphisms and Functional Variation". Drug Metab Dispos 30(10): 1043-1052.

Cashman, J.R and Zhang, J (2006). "Human Flavin-containing Monooxygenases". Ann Rev Pharmacol Toxicol 46: 65-100.

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

Charlesworth, D and Charlesworth, B 1987. "Inbreeding Depression and its Evolutionary Consequences". Ann Rev Ecol Syst 18: 237-268.

Choi, H.S, Ki, J.K, Cho, E.H, Kim, Y.C, Kim, J.I and Kim, S.W (2003). "A Novel Flavin-containing Monooxygenase from *Methylphaga* sp Strain SK1 and its Indigo Synthesis in *Escherichia coli*". Biochem Biophys Res Commun 306(4): 930-936.

Cherrington, N.J, Falls, J.G, Rose, R.L, Clements, K.M, Philpot, R.M, Levi, P.E and Hodgson, E (1998). "Molecular Cloning, Sequence, and Expression of Mouse Flavin-containing Monooxygenases 1 and 5 (*FMO1* and *FMO5*)". J Biochem Mol Toxicol 12(4): 205-212.

Clark, J.D (1984). From Hunters to Farmers: the Causes and Consequences of Food Production in Africa, University of California Press.

Collins, R.O (1971). Land Beyond the Rivers: the Southern Sudan, 1898-1918. Yale University Press. New haven**:** 53-57.

Conrad, D.F and Hurles, M.E (2007). "The Population Genetics of Structural Variation". Nat Genet 39(7 Suppl): S30-36.

Crentsil, P (2007). Death, Ancestors and HIV/Aids Among the Akan of Ghana. Sociology. University of Helsinki, Helsinki. PhD Thesis.

Diller, K.C, Gilbert, W.A and Kocher, T.D (2002). "Selective Sweeps in the Human Genome: A Starting Point for Identifying Genetic Differences Between Modern Humans and Chimpanzees". Mol Biol Evol 19(12): 2342-2345.

Dolphin, C, Cullingford, T.E, Shephard,E.A, Smith R.L and Phillips, I.R (1996). "Differential Development and Tissue-Specific Regulation of Expression of the

Genes Encoding Three Members of the Flavin-containing Monooxygenase Family of Man, *FMO1*, *FMO3* and *FMO4*". Eur J Biochem 235(3): 683-689.

Dolphin, C.T, Shephard, E.A, Povey, S, Palmer, C.N, Ziegler D.M, Ayesh, R, Smith R.L and Phillips, I.R (1991). "Cloning, Primary Sequence and Chromosomal Mapping of a Human Flavin-containing Monooxygenase (*FMO1*)". J Biol Chem 266(19): 12379-12385.

Dolphin, C.T, Beckett, D.J, Janmohamed, A, Cullingford, T.E, Smith, R.L, Shephard and Phillips, I.R (1998). "The Flavin-containing Monooxygenase 2 Gene (*FMO2*) of Humans, but not of Other Primates, Encodes a Truncated, Nonfunctional Protein". J Biolog Chem 273(46): 30599-30607.

Dolphin, C.T, Janmohamed, A, Smith, R.L, Shephard, E.A and Phillips, I.R (1997). "Missense Mutation in Flavin-containing Monooxygenase 3 Gene, *FMO3*, Underlies Fish-Odour Syndrome". Nat Genet 17(4): 491-494.

Dolphin, C.T, Janmohamed, A, Smith, R.L, Shephard, E.A and Phillips, I.R (2000). "Compound Heterozygosity for Missense Mutations in the Flavin-containing Monooxygenase 3 (*FMO3*) Gene in Patients with Fish-Odour Syndrome". Pharmacogenet 10(9): 799-807.

Dolphin, C.T, Shephard, E.A, Povey, S, Smith, R.L and Phillips, I.R (1992). "Cloning, Primary Sequence and Chromosomal Localization of Human *FMO2*, a New Member of the Flavin-containing Monooxygenase Family". Biochem J 287(Pt 1): 261-267.

Duescher, R.J, Lawton, M.P, Philpot, R.M and El Faro, A.A (1994). "Flavin-containing Monooxygenases (FMO)-Dependent Metabolism of Methionine Sulfoxidation in Rabbit Liver and Kidney". J Biol Chem 269(26): 17525-17530.

Duggan, D.E, Hooke, K.F, Risley, E.A, Shen, T.Y and Arman, C.G (1977). "Identification of the Biologically Active Form of Sulindac". J Pharmacol Exp Thr 201(1): 8-13.

Ehret, C. (2002). The Civilization of Africa, University of Virginia Press.

Eswaramoorthy, S, Bonanno, J.B, Burley, S.K and Swaminathan, S (2006). "Mechanism of Action of a Flavin-containing Monooxygenase". PNAS 103(26): 9832-9837.

Etienne, F, Resnick, L, Sagher, D, Brot, N and Weissbach, H (2003). "Reduction of Sulindac to its Active Metabolite, Sulindac Sulphide: Assay and Role of the Methionine Sulfoxide Reductase System". Biochem Biophys Res Commun. 312(4): 1005-1010.

Evans-Pritchard, E.E (1940). The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Oxford University Press. Oxford.

Excoffier, L, Laval G and Balding, D (2003). "Gametic Phase Estimation over Large Genomic Regions Using an Adaptive Window Approach". Hum Genom 1(1): 7-19.

Excoffier, L, Smousse, P.E and Quattro, J.M (1992). "Analysis of Molecular Variance Inferred from Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data". Genet 131(2): 479-491.

Excoffier, L and Slatkin, M (1995). "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population". Mol Biol Evol 12(5): 921-927.

Fay, J.C and Wu, C.I (2000). "Hitchhiking Under Positive Darwinian Selection". Genet 155(3): 1405-1413.

Fisher, M.B and Rettie A.E (1997). "Prochiral Sulfide Probes for the Active-Site Topography of Rabbit Flavin-containing Monooxygenase 2 (FMO2)". Tetrahedron Asymm 8(4): 613-618.

Fisher, R.A (1922). "On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P". J Royal Stat Soc 85(1): 87-94.

Fitch, W.M (1971). "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology". Syst Zoolog 20: 406-416.

Forster, P and Matsumina, S (2005). "Evolution. Did Early Humans go North or South?" Science 308(5724): 965-966.

Francois, A.A, Nishida, C.R, de Montellano, P.R.O, Phillips, I.R and Shephard, E.A (2009). "Human Flavin-containing Monooxygenase 2.1 Catalyzes Oxygenation of the Antitubercular Drugs Thiacetazone and Ethionamide". Drug Metab Dispos 37(1): 178-186.

Fu, Y and Li, W (1993). "Statistical Tests of Neutrality of Mutations". Genet 133(3): 693-709.

Fu, Y. X (1997). "Statistical Tests of Neutrality Against Population Growth, Hitchhiking and Background Selection". Genet 147(2): 915-925.

Fujieda, M, Yamazaki, H, Togashi, M, Saito, T and Kamataki, T (2003). "Two Novel Single Nucleotide Polymorphisms (SNPs) of the *FMO3* Gene in Japanese." Drug Metab Pharmacokinet 18(5): 333-335.

Fullerton, S.M, Clark, A.G, Weiss, KM, Nickerson, D.A, Taylor, S.L Stengard J.H, Salomaa, V, Vartiainen, E, Perola, M, Boerwinkle, E and Sing, C.F (2002). "Sequence Polymorphism at the Human Apolipoprotein A2 Gene (*APOA2*): Unexpected Deficit of Variation in an African-American Sample". Hum Gen 111:75-87.

Furnes, B.J, Feng, J, Sommer, S.S and Schlenk, D. (2003). "Identification of Novel Variants of the Flavin-containing Monooxygenase Gene Family in African Americans". Drug Metab Dispos 31(2): 187-193.

Furnes, B and Schlenk, D (2004). "Evaluation of Xenobiotic *N*- and *S*-Oxidation by Variant Flavin-containing Monooxygenase 1 (FMO1) Enzymes". Toxicol Sci. 78(2): 196-203.

Garrigan, D, Kingan, S.B, Pilkington, M.W, Wilder, J.A and Cox, M.P, Soodyall, H, Strassmann, B, Destro-Bisol, G, de Knijff, P, Novelletto, A, *et al* (2007). "Inferring Human Population Sizes, Divergence Times and Rates of Gene Flow from Mitochondrial, X and Y Chromosomes Resequencing Data". Genet 177: (4) 2195-2207.

Garrigan, D, Mobasher, Z, Severson, T, Wilder, J.A and Hammer, M.F (2004). "Evidence for Archaic Asian Ancestry on the Human X Chromosome". Mol Biol Evol 22(2): 189-192.

Gilks W.R, Richardson S, Spiegelhalter D.J (eds) (1996). Markov Chain-Monte Carlo in Practice. Chapman & Hall, London

Green, R.E, Krause, J, Briggs, A.W, Maricic, T, Stenzel, U, Kircher, M, Patterson, N, Li, H, Zhai, W, Fritz, M.H.Y *et al* (2010). " A Draft Sequence of the Neanderthal Genome". Science 328: 710-722.

Greenberg, J.H (1949). "Studies in African Linguistic Classification". Southwest J Anthropol 1(5): 79-100.

Griffiths, R. C and S. Tavare (1997). "Computational Methods for the Coalescent". IMA Volum Math Applic 87(10): 165-182.

Group, T. W. B (2010). World Development Indicators database from ddp-ext.worldbank.org/ext/ddpreport.

Guan, S, Falick, A.M, Williams, D.E and Cashman, J.R (1991). "Evidence for Complex Formation Between Rabbit Lung Flavin-containing Monooxygenase and Calreticulin". Biochem J 30(41): 9892-9900.

Guthery, S. L, Salisbury, B.A, Pungliya, M.S, Stephens, J.C and Bamshad, M (2007). "The Structure of Common Genetic Variation in United States Populations". Am J Hum Genet 81(6): 1221-1231.

Hamblin, M. T and Di Rienzo, A (2000). "Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus". Am J Hum Genet 66(5): 1669-1679.

Hamman, M. A, Haehner-Daniels, BD, Wrighton, S.A, Rettie A.E and Hall, S.D (2000). "Stereoselective Sulfoxidation of Sulindac Sulphide by Flavin-containing Monooxygenases". Biochem Pharmacol 60(1): 7-17.

Hao, D. C, Chen, S.L, Mu, J and Xiao, P.G (2009). "Molecular Phylogeny, Long-Term Evolution, and Functional Divergence of Flavin-containing Monooxygenases". Genetica 137(2): 173-187.

Harding, R.M, Fullerton, S.M, Griffith, R.C, Bond, J, Cox, M.J, Schneider, JA, Moulin, D.S and Clegg, J.B (1997). "Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans". Am J Hum Genet 60(4): 755-757.

Harpending, H.C (1994). "Signature of Ancient Population Growth in a Low-Resolution Mitochondrial DNA Mismatch Distribution". Hum Biol 66(4): 591-600.

Harris, E.E and Meyer, D (2006). "The Molecular Signature of Selection Underlying Human Adaptations". Am J Phy Anthropol 131(S43): 89-130.

Hawks, J, Wang, E.T, Cochran, G.M, Harpending, H.C and Moyzis, R.K (2007). "Recent Acceleration of Human Adaptive Evolution". Proc Natl Acad Sci USA 104(52): 20753-20758.

Hedrick, P. W (2005). "A Standardized Genetic Differentiation Measure". Evol 59(8): 1633-1638.

Hemminki, K, Rajaniemi, H, Lindahl, B and Moberger, B (1996). "Tamoxifen-induced DNA Adducts in Endometrial Samples from Breast Cancer Patients". Cancer Res 56: 4374-4377.

Henderson, M. C, Krueger S.K, Stevens J.F and Williams D.E. (2004). "Human Flavin-containing Monoxygenase Form 2 *S*-Oxygenation: Sulfenic Acid Formation from Thioureas and Oxidation of Glutathione". Chem Res Toxicol 17(5): 633-640.

Henderson, M.C, Siddons, L.K, Morre, J.T, Krueger, S.K and Williams, D.E (2008). "Metabolism of the Anti-Tuberculosis Drug Ethionamide by Mouse and Human FMO1, FMO2 and FMO3 and Mouse and Human Lung Microsomes." Toxicol Appl Pharmacol 15(233): 420-427.

Hernandez, D, Addou, S, Lee, D, Orengo, C, Shephard, E.A and Phillips, I.R (2003). "Trimethylaminuria and a Human *FMO3* Mutation Database". Hum Mutat 22(3): 209-213.

Hernandez, D, Janmohamed, A, Chandan, P, Phillips, I.R and Shephard E.A (2004). "Organization and Evolution of the Flavin-containing Monooxygenase Genes of Human and Mouse: Identification of Novel Gene and Pseudogene Clusters". Pharmacogenet 14(2): 117-130.

Henn, B.M, Gignoux, C.R, Jobin, M, Granka, J.M, Macpherson, J.M, Kidd, J.M, Rodriguez-Botigue, L, Ramachandran, S, Hon, L, Brisbin, A *et al.* (2011). "Hunter-gatherer Genomic Diversity Suggests a Southern African Origin for Modern Humans". PNAS Genet doi 10.1073

Hickey, A.J (2007). Lung Biology in Health and Disease: Inhalation Aerosols: Physical and Biological Basis for Therapy. Volume 94. Claude Lenfant, Editor. Informa Healthcare USA, inc.

Hill, W.G and Robertson, A (1968). "Linkage Disequilibrium in Finite Populations". Theor Appl Genet 38(6): 226-231.

Hiller, R (2007). The Principle of TaqMan Assays from http://www.cpgr.org.za/items/1171025441-0674.pdf.

Hines, R.N (2006). "Developmental and Tissue-Specific Expression of Human Flavin-containing Monooxygenases 1 and 3". Expert Opin. Drug Metab. Toxicol. 2(1): 41-49.

Hines, R.N, Hopp, K.A, Franco, J, Saeian, K and Begun, F.P (2002). "Alternative Processing of the Human *FMO6* Gene Renders Transcripts Incapable of Encoding a Functional Flavin-containing Monooxygenase". Mol Pharmacol 62(2): 320-325.

Hollox, E.J, Poulter, M, Zvarik, M, Ferak, V and Krause, A, Jenkins, T, Saha, N, Kozlov, A.L and Swallow, D.M (2001). "Lactase Haplotype Diversity in the Old World". Am J Hum Genet 68(1): 160-172.

Hudson, R.R and Kaplan, N. L (1985). "Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences". Genet 111(1): 147-164.

Hudson, R.S, Boos, D.D and Kaplan, N.L (1992). "A Statistical Test for Detecting Geographic Subdivision". Mol Biol Evol. 9(1): 138-152.

Hukkanen, J and Dempsey, D (2005). "Effect of Pregnancy on a Measure of FMO3 Activity". Br J Clin Pharmacol 60(2): 224-226.

International HapMap Consortium (2005). "A Haplotype Map of the Human Genome." Nature 437: 1299-1320.

Itagaki, K, Carver, G.T and Philpot R.M (1996). "Expression and Characterization of a Modified Flavin-containing Monooxygenase 4 from Humans". J Biol Chem 271(33): 20102-20107.

Itan, Y, Powell, A, Beaumont, M.A, Burger, J and Thomas, M.G (2009) "The Origins of Lactase Persistence in Europe". PLoS Computat Biolog 5(8):e1000491.

Jakobsson, M, Scholz, S.W, Scheet, P, Gibbs, J.R and VanLiere, J.M, Fung, H.C, Szpiech, Z.A, Degnan, J.H, Wang, K, Guerreiro, R, *et al* (2008). "Genotype, Haplotype and Copy-number Variation in Worldwide Human Populations." Nature 451: 998-1003.

Jenkins, O (1997). The Amhara People of Ethiopia. Retrieved 27th of June 2010 from http://orvillejenkins.com/profiles/amhara.html.

Jobling, M.A, Hurles, M and Tyler-Smith, C 2004. Human Evolutionary Genetics. Chapter 6. Garland Publishing. New York.

Johnston, M, Andrews, S, Brinkman, R, Cooper, J, Ding, H, Dover, J, Du, Z, Favello, A, Fulton, L, Gattung, S, *et al* (1994). "Complete Nucleotide Sequence of *Saccharomyces Cerevisiae* Chromosome VIII". Science 265(5181): 2077-2082.

Jost, L (2008). "G_{ST} and it`s Relative do not Measure Differentiation.". Molecul Ecolog 17(18): 4015-4026.

Jukes, T.H and Cantor, C.R (1969). Evolution of Protein Molecules. H.N. Munro, Editor. Mammalian Protein Metabolism 3. Academic Press. New York.

Kathamart, S and Stresser, D.M (2000). "Concurrent Flavin-containing Monooxygenase Down-Regulation and Cytochrome P-450 Induction by Dietary Indoles in Rat: Implications for Drug-Drug Interaction". Drug Metabol Dispos 28(8): 930-936.

Kelly, J.K (1997). "A Test of Neutrality Based on Interlocus Associations". Genet 146(3): 1197-1206.

Khomenko, T, Deng, X, Sandor, Z, Tarnawski, AS and Szabo, S (2004). "Cysteamine Alters Redox State, HIF-1 Alpha Transcriptional Interactions and Reduces Duedenal Ulceration: Novel Insight into the Mechanisms of Deudenal Ulcerations". Biochem Biophys Res Commun 317(1): 121-127.

Kim, U, Wooding, S, Ricci, D, Jorde, L.B and Drayna, D (2005). "Worldwide Haplotype Diversity and Coding Sequence Variation at Human Bitter Taste Receptor Loci". Hum Mutat 26(3): 199-204.

Kim, Y and Stephan, W (2000). "Joint Effects of Genetic Hitchhiking and Background Selection on Neutral Variation". Genet 155(3): 1415.

Kim, Y.H, Lee, J.Y, Lim, D.S, Shim, W.J, Ro, Y.M, Park G.H, Becker, K.G, Cho-Chung, Y.S and Kim, M.K (2003). "Gene Expression Profiling of Oxidative Stress on Atrial Fibrillation in Humans". Exp Mol Med 35(5): 336-349.

Kimura, M (1983). The Neutral Theory of Molecular Evolution. University of Cambridge. Cambridge, MA.

Kosakovsky, S.L, Frost, P.D.W and Frost, S.D.W (2005). "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection". Mol Biol Evol 22(5): 1208-1222.

Koukouritaki, S.B, Poch, M.T, Henderson, M.C, Siddens, L.K, Krueger, S.K, VanDyke, J.E, Williams, D.E , Pajewski, N.M, Wang, T and Hines, R.N (2007). "Identification and Functional Analysis of Common Human Flavin-containing Monooxygenase 3 Genetic Variants". J Pharmacol Exp Thr 320(1): 266-273.

Koukouritaki, S.B, Simpson, P, Yeung, C.K, Rettie, A.E and Hines, R.N (2002). "Human Hepatic Flavin-containing Monooxygenases 1 (FMO1) and 3 (FMO3) Developmental Expression". Pediatr Res 51(2): 236-243.

Krieter, P.A, Ziegler, D.M, Hill, K.E and Burk, R.F (1984). "Increased GSSG Efflux from Rat Livers Perfused with Thiocarbamides Substrates for the Flavin-containing Monooxygenase". Molec Pharmacol 26(1): 122-127.

Krueger, S.K, Siddens, L.K Henderson, M.C, Andreasen, E.A, Tanguay, R.L, Pereira, C.B, Cabacungan, E.T, Hines, R.N, Ardie, K.G and Williams, D.E (2005). "Haplotype and Functional Analysis of Four Flavin-containing Monooxygenase Isoform 2 (*FMO2*) Polymorphisms in Hispanics". Pharmacogenet genom 15(4): 245-256.

Krueger, S.K, Siddens, L.K, Martin, S.R, Yu, Z, Pereira, C.B, Cabacungan, E.T, Hines, R.N, Ardlie, K.G, Raucy, J.L and Williams, D.E (2004). "Differences in *FMO2* 1* Allelic Frequency between Hispanics of Puerto Rican and Mexican Descent". Drug Metab Dispos 32(12): 1337-1340.

Krueger, S.K and Williams, D.E (2005). "Mammalian Flavin-containing Monooxygenases: Structure/Function, Genetic Polymorphisms and Role in Drug Metabolism". Pharmacol Therapeut 106(3): 357-387.

Kuper, R and Kroepelin, S (2006). "Climate-Controlled Holocene Occupation in the Sahara: Motor of Africa`s Evolution". Science 313(5788): 803-807.

Kwiatkowski, D. P (2005). "How Malaria has Affected the Human Genome and What Human Genetics Can Teach Us About Malaria". Am J Hum Genet 77(2): 171-192.

Lambert, D.M, Mamer, O.A, Akerman, B.R, Choiniere, L, Gaudet, D and Hamet, P and Treacy, E.P (2001). "*In Vivo* Variability of TMA Oxidation is Partially Mediated by Polymorphism of the *FMO3* Gene". Mol Genet Metab 73(3): 224-229.

Lattard, V, Zhang, J, Tran, Q, Furnes, B , Schlenk, D and Cashman JR (2003). "Two New Polymorphisms of the *FMO3* Gene in Caucasians and African-American Populations: Comparative Genetic and Functional Studies". Drug Metab Dispos 31(7): 854-860.

Lawton, M.P, Cashman, J.R, Cresteil T, Dolphin C.T, Elfarra A.A, Hines R.N, Hodgson, E, Kimura, T, Ozols, J, Phillips, I.R, *et al*. (1994). "A Nomenclature for the Mammalian Flavin-containing Monooxygenase Gene Family Based on Amino Acid Sequence Identities". Arch Biochem Biophys 308(1): 254-257.

Lawton, M.P, Gasser, R, Tynes, R.E, Hodgson, E and Philpot, R.M (1990). "The Flavin-containing Monooxygenase Enzymes Expressed in Rabbit Liver and Lung are Products of Related but Distinctly Different Gene". J Biol Chem 265(10): 5855-5861.

Lee, M.Y, Clark, J.E and Williams D.E (1993). "Induction of Flavin-containing Monooxygenase (FMOB) in Rabbit Lung and Kidney by Sex Steroids and Glucocorticoids". Arch Biochem Biophys 302(2): 332-336.

Lewontin, R.C (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models". Genet 49(1): 49-67.

Librado, P and Rozas, J (2009). "DnaSP v5: A Software for Comprehensive Analysis of DNA Polymorphism Data". Bioinform 25:1451-1452.

Lin, J and Cashman, J.R (1997a). "Detoxification of Tyramine by the Flavin-containing Monooxygenase: Stereoselective Formation of the Trans Oxime." Chem Res Toxicol 10(8): 842-852.

Lin, J and Cashman, J.R (1997b). "*N*-Oxygenation of Phenethylamine to the Trans-Oxime by Adult Human Liver Flavin-containing Monooxygenase and Retroreduction of Phenethylamine Hydroxylamine by Human Liver Microsomes". J Pharmacol Exp Thr 282(3): 1269-1279.

Liu, H, Prugnolle, F, Manica, A and Balloux, F (2006). "A Geographically Explicit Genetic Model of Worldwide Human-Settlement History". Am J Hum Genet 79(2): 230-237.

Lomri, N, Gu, Q and Cashman, J.R (1992). "Molecular Cloning of the Flavin-containing Monooxygenase (form 2) cDNA from Adult Human Liver". Proc. Natl. Acad. Sci 89(5): 1685-1689.

Lomri, N, Yang, Z and Cashman J.R (1993). "Regio-and Stereoselective Oxygenations by Adult Human Liver Flavin-containing Monooxygenase 3. Comparison with Forms 1 and 2". Chem Res Toxicol 6(6): 800-807.

Macaulay, V, Hill, C, Achilli, A, Rengo, C, Clarke, D, Meeha, W, Blackburn, J, Semino, O, Scozzari, R, Cruciani, F, *et al.* (2005). "Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes". Science 308(5724): 1034-1036.

Malaspina, A, Kaushik, N and de Belleroche, J (2001). "Differential Expression of 14 Genes in Amyotrophic Lateral Sclerosis Spinal Cord Detected Using Gridded cDNA Arrays". J Neurochem 77(1): 132-145.

Massey, V (1994). "Activation of Molecular Oxygen by Flavins and Flavoproteins." J Biol Chem 269(36): 22459-22462.

McDonald, J.H and Kreitman, M (1991). "Adaptive Protein Evolution at the *ADH* Locus in Drosophila". Nature 351(6328): 652-654.

McDougall, I, Brown, F.H and Friegle, J.G (2005). "Stratigraphic Placement and Age of Modern Humans from Kibish, Ethiopia". Nature 433(7027): 733-736.

McElwain, K.V, Estienne, M.J and Barb, C.R (1999). "Effect of Cysteamine Hydrochloride on Secretion of Growth Hormone in Male Swine". Life Sci 64(24): 2233-2238.

Miller, M.M, James, R.A, Richer, J.K, Gordon, D.F, Wood, W.M and Horwitz, K.B (1997). "Progesterone-Regulated Expression of Flavin-containing

Monooxygenase 5 by B-Isoform of the Progesterone Receptors: Implications for Tamoxifen Carcinogenicity". J Clin Endocrinol.Metabol 82(9): 2956-2961.

Mitchell, S.C (2005). "Trimethylaminuria (Fish-Odour Syndrome) and Oral Malodour". Oral Dis. 11(1): 10-13.

Mitchell, S.C, Zhang, A.Q, Barret T, Ayesh, R and Smith, R.L (1997). "Studies on the Discontinous *N*-oxidation of Trimethylamine Among Jordanian, Ecuadorian and New Guinean Populations". Pharmacog 7(1): 45-50.

Mitchell, S.C and Smith, R.L (2001). "Trimethylaminuria: the Fish Malodor Syndrome". Drug Metab Dispos 29(4 Pt2): 517-521.

Mitchell, S.C and Smith, R.L (2003). "Trimethylamine and Odorous Sweat." Inherit Metab Dis 26(4): 415-416.

Mittl, P.R and Schultz, G.E (1994). "Structure of Glutathione Reductase from *Echerichia Coli* at 1.86. A Resolution: Comparison with the Enzyme from Human Erythrocytes". Prot Science 3(5): 799-809.

Miyata, T, S, Miyazawa, S and Yasunaga, T (1979). "Two Types of Amino Acid Substitutions in Protein Evolution". J Molec Evol 12(3): 219-236.

Miyata, T and Yasunaga, T (1980). "Molecular Evolution of mRNA: a Method for Estimating Evolutionary Rates of Synonymous and Amino Acid Substitutions from Homologous Nucleotide Sequences and its Application". J Mol Evol. 16(1): 23-26.

Nagata, T, Williams, D.E and Ziegler, D.M (1990). "Substrate Specificities of Rabbit Lung and Porcine Liver Flavin-containing Monooxygenase: Differences Due to Substrate Size". Chem Res Toxicol 3(4): 372-376.

Narimatsu, S, Tachibana, M, Masubuchi, Y and Suzuki, T (1996). "Cytochrome P4502D and -2C Enzymes Catalyze the Oxidative *N*-Demethylation of the Parkinsonism-Inducing Substance 1-Methyl-4-Phenyl-1,2,3,6-Tetrahydropyridine in Rat Liver Microsomes". Chem Res Toxicol 9: 93-98.

Neal, R.A and Halpert., J (1982). "Toxicology of Thiono-Sulfur Compounds." Ann Rev Pharmacol Toxicol 22: 321-329.

Neba, A (1999). Modern Geography of the Republic of Cameroon. Neba Publishers. Bamenda.

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

Nei, M and T. Gojobori (1986). "Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions". Mol Biol Evol 3(5)**:** 418-426.

Nei, M and F. Tajima (1981). "Genetic Drift and Estimation of Effective Population Size." Genet (3): 625-640.

Nida, W (2005). Gurage Ethno-Historical Survey. Encyclopaedia Aethiopica. Harrassowitz.Wiesbaden. 2**:** 929-935.

Nielsen, R, Bustamante, C, Clark, A.G, Glanowski, S and Sackton, T.B (2005). "A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees". PLoS Biol 3(6):e170.

Nielsen, R, Hellmann, I, Hubisz, M, Bustamante, C and Clark, A.G (2007). "Recent and Ongoing Selection in the Human Genome". Nat Rev Genet 8857-868..

Nikbakht, K. N, Lawton, M.P and Philpot, R.M (1992). "Guinea Pig or Rabbit Lung Flavin-containing Monooxygenases with Distinct Mobilities in SDS-PAGE are Allelic Variants that Differ at Only Two Positions". Pharmacogenet 2(5): 207-216.

Ntara, S.J (1973). The History of the Chewa. Franz Steiner Velag GMBH. Wiesbaden.

Ohmi, N, Yoshida, H, Endo, H, Hasegawa, M, Akimoto, M and Higuchi, S (2003). "*S*-Oxygenation of *S*-Methyl-Esonarimod by Flavin-containing Monooxygenases in Human Liver Microsomes". Xenobiotica 33: 1221-1231.

Ohmiya, Y and Mehendale., H.M (1982). "Metabolism of Chlorpromazine by Pulmonary Microsomal Enzymes in the Rat and Rabbit." Biochem Pharmacol 31(2): 157-162.

Oleksyk, T.K, Smith, M.W and O`Brien, S.T 2010. "Genome-wide Scans for Footprints of Natural Selection". Phil Trans R Soc B 365:1537: 185-205.

Onderwater, R.C.A, Commandeur, J.N.M, Menge, W.M.P.B and Vermeulen, N.P.E (1999). "Activation of Microsomal Glutathione *S*-Transferase and Inhibition of Cytochrome P450 1A1 Activity as a Model System for Detecting Protein Alkylation by Thiourea-Containing Compounds in Rat Liver Microsomes". Chem Res Toxicol 12(5): 396-402.

Overby, L.H, Buckpitt, A.R, Lawton, M.P, Atta-Asafo, E, Schulze J and Philpot, R.M (1995). "Characterization of Flavin-containing Monooxygenase 5 (FMO5)

Cloned from Human and Guinea Pig: Evidence that the Unique Catalytic Properties of FMO5 are not Confined to the Rabbit Ortholog". Arch Biochem Biophys 317(1): 275-284.

Overby, L.H, Carver, G.C and Philpot, R.M (1997). "Quantification and Kinetic Properties of Hepatic Microsomal and Recombinant Flavin-containing Monooxygenase 3 and 5 from Humans". Chem Biol Interact 106(1): 29-45.

Park, C.S, Kang, J.H, Chung, W.G, Yi, H.G, Pie, J.E, Park, D.L, Hines, R.N, McCarver, D.G and Cha, Y.N (2002). "Ethnic Differences in Allelic Frequency of Two Flavin-containing Monooxygenase 3 (*FMO3*) Polymorphisms: Linkage and Effect on *in Vivo* and *in Vitro* FMO Activities". Pharmacogenet 12(1): 77-80.

Parte P and Kupfer, D (2005). "Oxidation of Tamoxifen by Human Flavin-containing Monooxygenase (FMO) 1 and FMO3 to Tamoxifen-*N*-Oxide and its Novel Reduction Back to Tamoxifen by Human Cytochromes P450 and Hemoglobin". Drug Metab Dispos 33(10):1146-1452.

Peloquin, C.A (1993). "Pharmacology of the Antimycobacterial Drugs." Medical Clinics of North America 77(6): 1253-1262.

Phillips, I.R, Dolphin, C.T, Clair, P, Hadley, M.R, Hutt, A.J, McCombie, R.R, Smith, RL and Shephard, E.A (1995). "The Molecular Biology of the Flavin-containing Monooxygenases of Man". Chemico-Biolog Inter 96(1): 17-32.

Phillips, I.R, Francois, A,A and Shephard, E.A (2007). "The Flavin-containing Monoooxygenases (*FMO*s): Genetic Variation and its Consequences for the Metabolism of Therapeutic Drugs." Curr Pharmacogenom 5(4): 292-313.

Plackett, R.L (1983). "Karl Pearson and the Chi-Squared Test". Internat stat Rev 51(1): 59-72.

Plagnol, V and Wall J.D (2006) "Possible Ancestral Structure in Human Populations." PLoS Genet 2(7):e105.

Polyzos, A.A (2003). Directed Evolution of a Sulfoxidation Biocatalyst, Universty of Florida. PhD Thesis.

Poulsen, L.L (1981). "Organic Sulfur Substrates for the Microsomal Flavin-containing Monooxygenase". Rev Biochem Toxicol 3: 33-49.

Poulsen, L.L and Ziegler, D.M (1995). "Multisubstrate Flavin-containing Monooxygenase: Application of Mechanism to Specificity". Chem Biol Interact 96(1): 57-73.

Poulsen, L.L and Ziegler., D.M (1977). "Microsomal Mixed-Function Oxidase-Dependent Renaturation of Reduced Ribunuclease". Arch Biochem Biophys 183(2): 565-570.

Poulsen, L.L and Ziegler, D.M (1979). "The Liver Microsomal FAD-Containing Monooxygenase. Spectral Characterization and Kinetic Studies." J Biol Chem 254(14): 6449-6455.

Preacher, K.J (2001). Calculation for the Chi-Square Test: An Interactive Calculation Tool for Chi-Square Tests of Goodness of Fit and Independence. Computer Software.

Przeworski, M (2003). "Estimating the Time Since the Fixation of a Beneficial Allele". Genet 164: 1667-1676.

Quintana-Murci, L, Harmant, C, Quach, H, Blanovsky, O, Zaporozhchenko, V, Bormans, C, van Helden, P.D, Hoal, E.G amd Behar, D.M (2010) "Strong Maternal Khoisan Contribution to the South African Coloured Population: a Case of Gender-Biased Admixture". Am J Hum Genet 86(4): 611-620.

Quintana-Murci, L, Semino, O, Bandelt, H.J, Passarino, G, McElreavey, K and Santachiara-Benerecetti, A.S (1999). "Genetic Evidence of an Early Exit of *Homo Sapiens Sapiens* from Africa Through Eastern Africa". Nat Genet 23(4): 437-441.

Ramachandran, S, Deshpande, O, Roseman, C.C, Rosenberg, N.A, Feldman, M.W and Cavalli-Sforza, L.L (2005). "Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa". Proc Natl Acad Sci USA 102(44): 15942-15947.

Ramos-Onsins, S.E and Rochas, J (2002). "Statistical Properties of New Neutrality Tests Against Population Growth". Molec Biol Evol 19(12): 2092-2100.

Reader, J (1999). Africa: a Biography of the Continent. Vintage Books, New York.

Rebhan, M, Chalifa-Caspi, M, Prilusky, V and Lancet, J (1997). GeneCards: Encyclopedia for Genes, Proteins and Diseases, Weizmann Institute of World Wide Web

Reed, F.A and Tishkoff, S.A (2006). "African Human Diversity, Origins and Migrations". Curr Opin Genet Dev. 16(6): 597-605.

Rehfish, F (1960). "The Dynamics of Multilineality on the Mambila Plateau". J Internat Afr Instit 30(3): 246-261.

Reich, D, Green, R.E, Kircher, M, Krause, J, Patterson, N, Durand, E.Y, Viola, B, Briggs, A.W, Stenzel, U, Johnson, P.L.F *et al* (2010). "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia". Nature 468: 1053-1060.

Rettie, A.E, Lawton, M.P, Sadeque, A.J.M, Meier, G.P and Philpot (1994). "Prochiral Sulfoxidation as a Probe for Multiple Forms of the Microsomal Flavin-containing Monooxygenase: Studies with Rabbit FMO1, FMO2, FMO3 and FMO5 Expressed in *Escherichia Coli*". Arc Biochem Biophysics 311(2): 369-377.

Reynolds, J, Weir, B.S and Cockerham, C.C (1983). "Estimation of the Coancestry Coefficient Basis for a Short-Term Genetic Distance". Genet 105(3): 767-779.

Rodriguez, S, Gaunt, T.R and Ian, N.M.D (2009). "Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies. Am J Epidemiol 169(4):505-514.

Rogers, A.R and Harpending, H (1992). "Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences". Molec Biol Evol 9(3): 552-569.

Rosenberg, N.A, Pritchard, J.K, Weber, J.L, Cann, H.M, Kidd, K.K, Zhivotovsky, L.A and Feldman, M.W (2002). "Genetic Structure of Human Populations". Science 298: 2381-2385.

Rozas, J, Gullaud, M, Blandin, G and Aguade, M (2001). "DNA Variation at the RP49 Gene Region of *Drosophila Simulans:* Evolutionary Inferences from an Unusual Haplotype Structure". Genet 158: 1147-1155.

Sabbagh, A and Darlu, P (2005) "Inferring Haplotypes at the *NAT2* Locus: the Computational Approach". BMC Genet 6 :30 doi:10.1186/1471-2156-6-30.

Sabeti, P.C, Reich, D,E, Higgins, J.M, Levine, H.Z.P, Richter, D.J, Schaffner, S.F, Gabriel, S.B, Platko, J.V, Patterson, N.J and McDonald, G.J. (2002). "Detecting Recent Positive Selection in the Human Genome from Haplotype Structure". Nature 419(6909): 832-837.

Sabeti, P.C, Varilly, P, Fry, B, Lohmueller, J, Hostetter, E, Cotsapas, C, Xie, X, Byrne, E.H, McCarroll, S.A and Gaudet, R (2007). "Genome-wide Detection and Characterization of Positive Selection in Human Populations". Nature 449(7164): 913-918.

Sachidanandam, R, Weissman, D, Schmidt, S.C, Kakol, S.C, Stein, L.D, Marth, G, Sherry, s, Mullikin, J.C, Mortimore, B.J and Wiley, D.L (2001). "A Map of

Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms". Nature 409(6822): 928-933.

Sardas, S, Akyol, D, Green, R.L, Mellon, T, Gokmen, O and Cholerton, S (1996). "Trimethylamine *N*-oxidation in Turkish Women with Bacterial Vaginosis." Pharmacogenet 6(5): 459-463.

Schaffner, S.F and Sabeti P. (2008). "Evolutionary Adaptation in the Human Lineage". Nature Education 1(1).

Schott, R (1977). "Sources for a History of the Bulsa in Northern Ghana." Paideuma 23: 141-168.

Schott, R (1987). "Traditional Law and Religion Among the Bulsa of Northern Ghana" J Afr Law 31(1-2): 58-69.

Shack, W (1966). The Gurage. A People of the Ensete Culture. Oford University Press. Oxford.

Shephard, E.A, Chandan, P, Stevanovic-Walker, M, Edwards, M and Phillips, I.R (2007). "Alternative Promoters and Repetitive DNA Elements Define the Species-Dependent Tissue-Specific Expression of the *FMO1* Genes of Human and Mouse". Biochem J 406(3): 491-499.

Shimizu, M, Fujita, H, Aoyama, T and Yamazaki, H (2006). "Three Novel Single Nucleotide Polymorphisms of the *FMO3* Gene in a Japanese Population". Drug Metab Pharmacokinet 21(3): 245-247.

Shinn, D.H, Ofcansky, T.P and Prouty C (2004). Historical Dictionary of Ethiopia. Lanham. The Scarecrow Press

Sirugo, G, Hennig, B.J, Adeyemo, A, Matimba, A and Newport M.J (2008). "Genetic Studies of African Populations: an Overview on Disease Susceptibility and Response to Vaccines and Therapeutics". Hum Genet 123(6): 557-598.

Maynard-Smith, J and Haigh, J (1974). " The Hitch-hiking Effect of a Favourable Gene". Genet Res 23: 23-35.

Smith, P.B and Crespi, C (2002). "Thiourea Toxicity in Mouse C3H/10T1/2 Cells Expressing Human Flavin-Dependent Monooxygenases 3". Biochem Pharmacol 63(11): 1941-1948.

Soriano, A.R (2008). "Selection and Linkage Disequilibrium Tests Under Complex Demographies and Ascertainment Bias. University of Barcelona. Barcelona. PhD Thesis.

Stahl, E.A and Bishop, J.G (2000). "Plant-pathogen Arms Races at the Molecular Level". Curr Opin Plant Biol 3(4): 299-304.

Stehr, M, Diekmann, H, Smau, L, Seth, O, Ghisla, S, Singh, M and Macheroux, P (1998). "A Hydrophobic Sequence Motif Common to *N*-Hydroxylating Enzymes". Trends Biochem Scienc 23(2): 56-57.

Stephens, M, Smith N.J and Donnelly, P (2001). "A New Statistical Method for Haplotype Reconstruction from Population Data". Am J Hum Genet 68(4): 978-989.

Stephens, M and Donnelly, P (2003). "A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data". Am J Hum Gen. 73(5): 1162-1169.

Storz, J.F, Payseur, B.A and Nachman, M.W (2004). "Genome Scans of DNA Variability in Humans Reveal Evidence for Selective Sweeps Outside of Africa". Mol Biol Evol 21(9): 1800-1811.

Stringer, C.B and Andrews, P (1988). "Genetic and Fossil Evidence for the Origin of Modern Humans". Science 239(4845): 1263-1268.

Stringer, C.B and Andrews, P (2005). The Complete World of Human Evolution. Thames and Hudson.

Suh, J.K, Poulsen, L.L, Ziegler, D.M and Robertus, J.D (1996). "Molecular Cloning and Kinetic Characterization of a Flavin-containing Mooxygenase from *Saccharomyces Cerevisiae*". Arch Biochem Biophys 336(2): 268-274.

Tajima, F (1989). "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism". Genet 123(3): 585-595.

Takamura, M, Sakurai, M, Ota, T, Ando, H, Honda, M and Kaneko, S (2004). "Genes for Systemic Vascular Complications are Differentially Expressed in the Livers of Type 2 Diabetic Patients". Diabetologia 47(4): 638-647.

Thangaraj, K, Chaubey, G, Kivisild, T, Reddy, A.G and Singh, V.K, Rasalkar, A.A and Singh, L (2005). "Reconstructing the Origin of Andaman Islanders". Science 308(5724): 996.

Tishkoff, S.A, Dietzsch, E, Speed, W, Pakstis, A.J, Kidd, J.R, Cheung, K, Bonnè-Tamir, B, Santachiara-Benerecetti, A.S, Moral, P and Krings, M (1996). "Global Patterns of Linkage Disequilibrium at the *CD4* Locus and Modern Human Origins". Science 271(5254): 1380-1387.

Tishkoff, S.A, Gonder, M.K, Henn, B.M, Mortensen, H, Knight, A, Gignoux, C, Fernandopulle, N, Lema,G, Nyambo, T.B, Ramakrishnan, U, *et al.* (2007a). "History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation". Mol Biol Evol 24(10): 2180-2195.

Tishkoff, S.A, Reed, F.A, Friedlander, F.R, Ehret, C, Ranciaro, A, Froment, A, Hirbo, J.B, Awomoyi, A.A, Bodo, J.M, Doumbo, O, *et al.* (2009). "The Genetic Structure and History of Africans and African Americans". Science 324(5930): 1035-1044.

Tishkoff, S.A, Reed, F.A, Ranciaro, A, Voight, B.F, Babbitt, C.C, Silverman, J.S, Powell, K, Mortensen, H.M, Hirbo, J.B, Osman, M, *et al.* (2007b). "Convergent Adaptation of Human Lactase Persistence in Africa and Europe". Nat. Genet 39(1): 31-40.

Tishkoff, S.A, Varkonyi, R, Cahinhinan, N, Abbes, S, Argyropoulos, G, Destro-Bisol, G, Drousiotou, A, Dangerfield, B, Lefranc, G, Loiselet, J. *et al.* (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". Ann Rev Genom Hum Genet 4(1): 293-340.

Tishkoff, S.A and Williams, S.M (2002). "Genetic Analysis of African Populations: Human Evolution and Complex Disease". Nat Rev Genet 3(8): 611-621.

Treacy, E.P, Akerman, B.R, Chow, L.M.L, Youil, R, Bibeau, C and Lin, J (1998). "Mutations of the Flavin-containing Monooxygenase Gene (*FMO3*) Cause Trimethylaminuria, a Defect in Detoxification". Hum Mol Genet 7(5): 839-845.

Vanelli, T.A, Dykman, A and Ortiz de Montellano, P.R (2002). "The Antituberculosis Drug Ethionamide is Activated by a Flavoprotein Monooxygenase". J Biol Chem. 277:12824-12829.

Vansina, J (1995). "New Linguistic Evidence and the Bantu Expansion". J Afr Hist 36(2): 173-195.

Veeramah, K. R, Thomas, M. G, Weale, M.E, Zeitlyn, D, Tarekegn, A, Bekele, E, Mendell, N.R, Shephard, E.A, Bradman, N and Phillips, I.R (2008). "The

Potentially Deleterious Functional Variant Flavin-containing Monooxygenase 2* 1 is at High Frequency Throughout sub-Saharan Africa". Pharmacogenet genom 18(10): 877-886.

Verrelli, B.C, McDonald, J.H, Argyropolous, G, Destro-Bisol, G, Froment, A, Drousiotou, A, Lefranc, G, Helal, A.N, Loiselet, J and Tishkoff, S.A (2002). "Evidence for Balancing Selection from Nucleotide Sequence Analyses of Human G6PD". Am J Hum Genet 71(5):1112-1128.

Voight, B.F, Kudaravalli, S, Wen, X and Pritchard, J.K. (2006) "A Map of Recent Positive Selection in the Human Genome". PLoS Biol 4(3):e72..

Von Cramon-Taubadel, N. and Lycett, S. J. (2008). "Brief Communication: Human Cranial Variation Fits Iterative Founder Effect Model with African Origin". Am J Phys Anthropol 136(1): 108-113.

Wall, J.D (1999). "Recombination and the Power of Statistical Tests of Neutrality". Genet Res 74(1): 65-79.

Wall, J.D (2002). "Estimating Ancestral Population Sizes and Divergence Times". Genet 163r: 395-404.

Wall, J.D, Lohmueller K.E and Plagnol, V (2009). "Detecting Ancient Admixture and Estimating Demographic Parameters in Multiple Human Populations". Mol Biol Evol 26(8): 1823-1827.

Wang, X, Grus, W.E and Zhang, J (2006). "Gene Losses During Human Origins". PLoS Biol 4(3): e52.

Watterson, G.A (1975). "On the Number of Segregating Sites in Genetical Models Without Recombination". Theoretic Pop Biol 7(2): 256-276.

Weekes, R.V (1984). "Afar", Muslim Peoples. Westport, Greenwood Press. Connecticut.

Weir, B, Cardon, L.R, Anderson, A.D, Nielsen, D.M and Hill, W.G (2005). "Measures of Human Population Structure Show Heterogeneity Among Genomic Regions". Genome Res. 15: 1468-1476.

Whetstine, J.R, Yueh, M.F, Hopp, K.A, McCarver, D.G, Williams, D.E, Park, C.S, Kang, J.H, Cha, Y.N, Dolphin, C.T and Shephard, E.A (2000). "Ethnic Differences in Human Flavin-containing Monooxygenase 2 (*FMO2*) Polymorphisms: Detection of Expressed Protein in African-Americans". Toxicol Appl Pharmacol 168(3): 216-224.

White, T, Asfaw, B, DeGusta, D, Gilbert, H, Richards, G.D, Suwa, G and Howell, F.C (2003). "Pleitocene *Homo Sapiens* from Middle Awash, Ethiopia". Nature 423(6941): 742-747.

Wieranga, R.K, De Maeyer, M.C.H and Hol, W.G.J (1985). "Interactions of Pyrophosphate Moities with A-Helixes in Dinucleotide Binding Proteins". Biochem 24(6): 1346-1357.

Williams, D.E, Ziegler D.M, Nordin, D.J, Hale, S.E and Masters B.S (1984). "Rabbit Lung Flavin-containing Monooxygenase is Immunochemically and Catalytically Distinct from the Liver Enzyme". Biochem Biophys Res Commun 125(1): 116-122.

Williamson, S.H, Hubisz, M.J, Clark, A.G, Payseur, B.A, Bustamante, C.D and Nielsen, R (2007) "Localizing Recent Adaptive Evolution in the Human Genome". PLoS Genet 3(6):e90.

Wilson, J.F, Weale, M.E, Smith, A.C, Gratrix, F.F, Flethcher, B, Thomas, M.G, Bradman, N and Goldstein, D.B (2001). "Population Genetic Structure of Variable Drug Response". Nat. Genet 29(3): 265-269.

Wolpoff, M.H, Hawks, J and Caspari, R (2000). "Multiregional, not Multiple Origins". Am J Phys Anthr 112: 129-136.

Wood, E.T, Stover, D.A, Ehret, C, Destro-Bisol, G, Spedini, G, McLeod, H, Louie, L, Bamshad, M, Strassmann, B.I, Soodyall. H, *et al* (2005). "Contrasting Patterns of Y Chromosome and mtDNA Variation in Africa: Evidence for Sex-Biased Demographic Processes". Eur J Hum Genet 13(7): 867-876.

Wooding, S.P, Watkins, W.S, Bamshad, M.J, Dunn, D.M, Weiss, R.B and Jorde, L.B (2002). "DNA Sequence Variation in a 3.7-kb Noncoding Sequence 5`of the *CYP1A2* Gene: Implications for Human Population History and Natural Selection". Am J Hum Genet. 71(3): 528-542.

Wright, S (1931). "Evolution in Mendelian Populations". Genet 16: 97-159.

Wu, R.F and Ichikawa, Y (1995). "Inhibition of 1-Methyl-4-Phenyl-1, 2, 3, 6-Tetrahydropyridine Metabolic Activity of Porcine FAD-Containing Monooxygenase by Selective monoamine oxidase-B Inhibitors". FEBS Lett 358(2): 145-148.

Wyatt, M.K, Philpot, R.M, Carver, G, Lawton, M.P and Nikbakht, K.N (1996). "Structural Characteristics of Flavin-containing Monooxygense Genes One and Two (*FMO1* and *FMO2*)". Drug Metab Dispos 24(12): 1320-1327.

Xi, T, Jones, I.M and Mohrenweiser, H.W (2004). "Many Amino Acid Substitution Variants Identified in DNA Repair Genes during Human Population Screenings are Predicted to Impact Protein Function". Genom 83(6): 970-979.

Yakan, M.A (1999). Almanac of African Peoples & Nations. Transaction Publishers.

Yeung, C.K, Lang, D.H, Thummel, K.E and Rettie, A.E (2000). "Immunoquantitation of FMO1 in Human Liver, Kidney and Intestine". Drug Metab Dispos 28(9): 1107-1111.

Yotova, V, Lefebvre, J.F, Kohany, O, Jurka, J, Michalski R, Modiano, D, Utermann, G, Williams, S.M and Labuda, D (2007). "Tracing Genetic History of Modern Humans using X-Chromosome Lineages". Hum Genet 122(5): 431-443.

Yueh, M.F, Krueger S.K and Williams, D.E (1997). "Pulmonary Flavin-containing Monooxygenase (FMO) in Rhesus Macaque: Expression of FMO2 Protein, mRNA and Analysis of the cDNA". Biochem Biophys Acta 1350(3): 267-271.

Zeitlyn, D (1994). Aspects of Mambila Traditional Religion. Academia Verlag. Sankt Augustin.

Zeng, K, Fu, Y.X, Shi, S and Wu, C.I (2006). "Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants". Genet 174(3): 1431-1439.

Zhang, J and Cashman, J. R (2006). "Quantitative Analysis of *FMO* Gene mRNA Levels in Human Tissues". Drug Metabol Dispos 34(1): 19-26.

Zhang, J, Tran, Q, Lattard, V and Cashman, J.R (2003). "Deleterious Mutations in the Flavin-containing Monooxygenase 3 (*FMO3*) Gene Causing Trimethylaminuria". Pharmacogenet 13(8): 495-500.

Zhang, J.Y, Wang, Y and Prakash, C (2006a). "Xenobiotic-Metabolizing Enzymes in Human Lung". Curr Drug Metab 7(8): 939-948.

Zhang, Z, Li, J, Zhao, X.Q, Wang, J, Wong, G.K.S and Yu, J (2006b). "Ka/Ks Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging". Genom Proteom Bioinform 4(4): 259-263.

Zhao, Z, Fu,Y.X, Hewett-Emmet, D and Boerwinkle, E (2003). "Investigating Single Nucleotide Polymorphism (SNP) Density in the Human Genome and its Implications for Molecular Evolution". Gene 312: 207-213.

Ziegler, D.M (2002). "An Overview of the Mechanism, Substrate Specificities and Structure of FMOs". Drug Metab Rev 34(3): 503-511.

Ziegler, D.M (1990). "The 1990 Bernard B Brodie Award Lecture. Unique Properties of the Enzymes of Detoxification". Drug Metabol Disposit 19: 847-852.

Ziegler, D.M (1988). "Flavin-containing Monooxygenases: Catalytic Mechanism and Substrate Specificities". Drug Metab Rev 19(1): 1-32.

Ziegler, D.M (1991). "Mechanism, Multiple Forms and Substrate Specificities of Flavin-containing Monooxygenases. *N*-Oxidation of Drugs: Biochemistry, Pharmacology and Toxicology. Chapman and Hall. London.

Ziegler, D.M (1993). "Recent Studies on the Structure and Function of Multisubstrate Flavin-containing Monooxygenases". Ann Rev Pharmacol Toxicol 33: 179-199.

Ziegler, D. M and Petit, F.H (1964). "Formation of an Intermediate *N*-Oxide in the Oxidative Demethylation of *N*, *N*-Dimethylaniline Catalyzed by Liver Microsomes" Biochem Biophys Res Commun 15: 188-193.

Zschocke, J, Kohlmueller D, Quak, E, Meissner, T, Hoffman, G.F and Mayatepek, E (1999). "Mild Trimethylaminuria Caused by Common Variants in *FMO3* Gene". Lancet 354(9181): 834-835.