Pharmacogenetics of African populations:

Variation in major drug metabolising enzyme genes and potential impact on personalised medicine

Alice Matimba

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Division of Human Genetics

University of Cape Town

August 2009

Supervisors

Professor Raj Ramesar

Dr. Collen Masimirembwa

We are hard-pressed on every side, yet not crushed; we are perplexed, but not in despair; persecuted, but not forsaken; struck down, but not destroyed

2 Corinthians 4; 8,9

TABLE OF CONTENTS

ACKNOWLEDGEMENTS vi
ABSTRACT vii
LIST OF ABBREVIATIONS ix
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF PUBLICATIONS xiii

1. INTRODUCTION 1

1.1. Human Genetics and Diversity 1
1.1.1. Inheritance: genetics and evolution 2
1.1.2. Phenotype and genotype 3
1.1.3. Variation in the human genome 4
1.1.4. Genetic diversity of populations 6

1.2. Africa 9
1.2.1. The African continent and its people 9
1.2.2. African languages and ethnicities 9
1.2.3. Disease burden in Africa 11

1.3. Personalised Medicine 14
1.3.1. Genomics-informed medicine 14
1.3.1.1. Monogenic and complex disorders 15
1.3.1.2. Genetic susceptibility to infection 16
1.3.1.3. Drug response as a complex trait 16
1.3.1.4. Genetic testing and prediction 17
1.3.2. Resources for genomic research 17
1.3.2.1. Methods in genomics 18
1.3.2.2. Biobanks 21
1.3.2.3. Genetic databases 24

1.4. Pharmacogenetics 26
1.4.1. Drug response and pharmacogenetics 26
1.4.1.1. History and current advances in pharmacogenetics 26
1.4.1.2. Properties of drug response 27
1.4.2. ADMET genes 28
1.4.3. Variation in drug metabolism 30
1.4.4. Drug metabolising enzymes 31
1.4.4.1. Cytochrome P450 (CYP) 31
1.4.4.1.1 CYP2B 33
1.4.4.1.2 CYP2C 35
1.4.4.1.3 CYP2D 36
1.4.4.2. Flavin mono-oxygenases (FMO) 37
1.4.4.3. Glutathione S-transferases (GST) 37
1.4.4.4. N-acetyltransferases (NAT) 39

1.4.5. Personalised medicine and pharmacogenomics 40
1.4.5.1. Pharmacogenomics in drug development 41
1.4.5.2. Pharmacodiagnostic markers 41

1.5. Building African capacity for the future 46
1.5.1. Bioresources for pharmacogenetics research 46
1.5.2. Pharmacogenomics knowledge in Africa 46

2. AIMS AND OBJECTIVES 48

2.1. AIM 48

2.2. OBJECTIVES 48

3. MATERIALS AND METHODS 49

3.1. Populations 49
3.1.1. Ethics 49
3.1.2. Population selection and volunteers 49
3.1.3. Blood samples 50
3.1.4. DNA extraction and quality control 50

3.2. Genes 50
3.2.1. Selected genes 50
3.2.2. Allele nomenclature 51

3.3. Re-sequencing 52

3.4. Genotyping 55
3.4.1. RFLP 55
3.4.2. Taqman 56
3.4.3. Sequenom 60

3.5. Collection of reference data 61

3.6. Data analysis 61
3.6.1. Nucleotide sequence 61
3.6.2. Haplotype determination 62
3.6.3. Prediction of functional effects of non-synonymous SNPs 62
3.6.4. Splice site recognition and mRNA processing 63
3.6.5. Statistical analysis 63
3.6.5.1. Population differentiation 63
3.6.5.2. Genetic and geographic distances 64
3.6.5.3. Principal Component Analysis (PCA) 66
3.6.5.4. Phylogenetic analysis 66
3.6.5.5. Analysis of Molecular Variance (AMOVA) 66

3.7. Data records **67**
3.7.1. Sample and genotype management 67
3.7.2. Cataloguing and ranking of polymorphisms 68

4. RESULTS **69**

4.1. Re-sequencing analysis of DME genes **70**
4.1.1. Characterisation of non-synonymous SNPs 82
4.1.2. Determination of splice variants 83
4.1.3. Haplotype determination 83

4.2. Analysis of HapMap SNPs **87**

4.3. Baseline prevalence of common alleles **89**

4.4. Population differentiation and relatedness **94**
4.4.1. Population variation in SNPs from re-sequencing analysis 94
4.4.2. Population variation in HapMap SNPs 97
4.4.3. Population variation in common alleles 99

4.5. Building pharmacogenetics resources in Africa **105**
4.5.1. Demographics of sample collection 105
4.5.2. Sample and genotype database 106
4.5.3. Catalogue of African polymorphisms 108
4.5.4. Pharmacodiagnostic kit 108

5. DISCUSSION **118**

5.1. Discovery of novel SNPs in DME genes **118**
5.1.1. Functional characterisation of mutations 120
5.1.2. SNPs affect mRNA processing 124
5.1.3. Validation of functional effects of SNPs 125

5.2. Complex haplotype networks **127**
5.2.1. Allele nomenclature and new alleles 127
5.2.2. HapMap SNPs 128

5.3. Prevalence and clinical impact of DME polymorphisms **130**
5.3.1. CYP2B6 131
5.3.2. CYP2C9 132
5.3.3. CYP2C19 133
5.3.4. CYP2D6 134
5.3.5. FMO3 136
5.3.6. GST 138
5.3.7. NAT2 138

5.4. Population differentiation and evolutionary relatedness **140**

5.4.1. Differentiation of African populations 140
5.4.2. Evolutionary relatedness based on common alleles 143
5.4.3. Pharmacogenetic variants as population markers 145

5.5. Africa: Bioresource building **147**
5.5.1. African populations 147
5.5.2. African pharmacogenetics 149
5.5.3. African biorepositories 154
5.5.4. African Databases 155
5.5.4.1. Sample and genotype database 155
5.5.4.2. Pharmacogenetics database 156

5.6. PGxA - Pharmacogenetics for Personalised Medicine in Africa **158**
5.6.1. Benefits for treatment and exposure 158
5.6.1.1. Minimising ADRs 158
5.6.1.2. Improving drug economy 159
5.6.1.3. Monitoring patient compliance 159
5.6.1.4. Responding to environmental toxins 160
5.6.2. Pharmacogenetics for personalised therapy 160
5.6.3. A pharmacodiagnostic kit for African populations 162

CONCLUSIONS **165**

REFERENCES **164**

APPENDIX

ACKNOWLEDGEMENTS

Collen Masimirembwa, who gave me the opportunity to research such an exciting topic, and for his supervision in planning and undertaking experimental work at the **African Institute of Biomedical Science and Technology (AiBST)**.

Flemish Interuniversity Council (VLIR) for funding the research. I also acknowledge the supervision of Dr. Juergen Del Favero and Prof. Dr. Christine Van Broeckhoven during my studies at the University of Antwerp, Belgium.

Raj Ramesar and the **Division of Human Genetics**, for advising me in the final stages of the work and enabling the incorporation of this thesis at the University of Cape Town, South Africa.

Marja-Liisa Dahl and **Mao Mao** for facilitating the FMO3 work at the University of Uppsala, Sweden.

This work would not have been possible without the individuals who donated their samples and the coordinated collection efforts by the **Consortium of Pharmacogenetics of African Populations (CoPhA)**. Special acknowledgements to Prof. Guantai, Prof. Bolaji, Prof. Sayi, Collet Dandara, Emmanuel Chigutsa, Margaret Oluka and Ben Ebeshi with whom I worked closely during my studies.

Above all, I thank my family, whose constant encouragement, love and support I have relied upon throughout. It is to them that I dedicate this work.

ABSTRACT

Individual drug response is a complex trait, shaped by a person's genetic profile. Pharmacogenetics aims at understanding its underlying variability, enabling personalised treatment with minimal adverse effects. Polymorphism of drug metabolising enzyme genes, such as cytochrome P450s (CYPs), flavin mono-oxygenase (FMO), glutathion S-transferase (GST) and N-acteyltransferases (NAT), has been documented in Asian and Caucasian populations, yet studies on African variants have been scarce. Therefore, here the most comprehensive analysis of drug metabolising enzyme genes in a diverse African population sample was undertaken.

Eight drug metabolising enzyme genes - *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *FMO3*, *GSTM1*, *GSTT1* and *NAT2* - were characterised in ten African populations - Hausa, Igbo, Yoruba, Kikuyu, Luo, Maasai, San, Shona, Venda and Tanzanian Bantu. Re-sequencing analysis was used to find novel variants and ascertain known polymorphisms. Their impact on mRNA processing and protein function was estimated, using bioinformatic applications. Restriction fragment length polymorphism (RFLP), Taqman and Sequenom high-throughput multiplex genotyping were employed to determine baseline frequencies of known alleles. Using statistical models, African population diversity and relatedness was explored.

Novel SNPs, which may affect enzyme function due to amino acid changes, were detected in *CYP2C9*, *CYP2C19*, *CYP2D6* and *NAT2*. A high prevalence of low frequency or rare variants was observed. Diversity in the frequencies of polymorphisms between African

ethnicities was generally low, with most variation occurring *within* populations.

Genotype data was recorded, and polymorphisms were ranked towards the development of a pharmacogenetics database and a pharmacodiagnostic kit for African populations.

LIST OF ABBREVIATIONS

A = Adenine
aa = amino acid
ABC = ATP-Binding Cassette
ADMET = Absorption Distribution Metabolism Excretion Toxicity
ADR = Adverse Drug Reaction
AIDS= Acquired Immunodeficiency Syndrome
AMOVA = Analysis of Molecular Variance
ANOVA = Analysis of Variance
ART = Anti-retroviral treatment
BLAST = Basic Local Alignment Search Tool
bp = base pairs
C = Cytosine
cDNA = complementary DNA
CNV = Copy Number Variant
CYP = Cytochrome P-450
dbSNP = Single Nucleotide Polymorphism database
del = deletion
DME = Drug Metabolising Enzyme
DNA = Deoxyribonucleic Acid
dNTP = deoxyribonucleotide triphosphate
DTC = Direct-To-Consumer
Dx = Companion Diagnostic
EH = Epoxide hydrolase
FDA = Food and Drug Administration
FMO = Flavin mono-oxygenase
fs = frameshift
Fst = Fixation index
G = Guanine
GWAS = Genome Wide Association Study
GST = Glutathione S-transferase
HIV = Human Immunodeficiency Virus
HPLC = High Performance Liquid Chromatography
HW = Hardy Weinberg equilibrium
indel = insertion/deletion
ins = insertion
kb = kilo bases
LD=Linkage Disequilibrium
LOD = Logarithm of Odds
MALDI-TOF = Matrix-Assisted Laser Desorption/Ionisation-Time-of-Flight
MDR = Multi-drug resistant
MDR1 = Multi-drug transporter 1
mRNA = messenger RNA

MT = Mutant Allele
MW = Molecular Weight
NAT = N-actetyltransferase
NCBI = National Center for Biotechnology Information
ns = non-synonymous
NSAID = Non-Steroidal Anti-Inflammatory Drug
OMIM = Online Mendelian Inheritance in Man
PCA = Principal Component Analysis
PCR = Polymerase Chain Reaction
PD = Pharmacodynamics
PDB = Protein Data Bank
PGx = Pharmacogenetics
PHYLIP = Phylogeny Inference Package
PK = Pharmacokinetics
premiRNA = pre-micro RNA
PSIC = Position-Specific Independent Count
RE = Restriction Enzymes
RFLP = Restriction Fragment Length Polymorphism
RMA = Reduced Major Axis
RNA = Ribonucleic Acid
s = synonymous
SNP = Single Nucleotide Polymorphism
STR = Short Tandem Repeat
SULT = Sulphonyl transferase
T = Thymine
TB = Tuberculosis
TPMT = Thiopurine S-methyl transferase
TZB = Tanzanian Bantu
UGT = Uridine glucuronyl transferase
UPGMA = Unweighted Pair Group Method Arithmetic
UTR = Untranslated Region
VKORC = Vitamin K epoxide reductase
WT = Wild type
XDR = Extensively drug-resistant
YRI = Yoruba

LIST OF TABLES

Table 1: Some ethnic groups from eastern, western and southern Africa

Table 2: Examples of Biorepositories

Table 3: Examples of pharmacogenetic applications in drug therapy

Table 4: Genes analysed in this study

Table 5: Exon amplification and sequencing primers

Table 6: Alleles, primers, PCR conditions, restriction enzymes (RE) and expected fragment pattern of RFLP analysis

Table 7: Primers for analysis using Sequenom MassARRAY

Table 8: *CYP2C9* SNP frequencies

Table 9: *CYP2C19* SNP frequencies

Table 10: *CYP2D6* SNP frequencies

Table 11: *NAT2* SNP frequencies

Table 12: Grouping of novel SNPs and functional effect prediction

Table 13: Genotype frequencies of HapMap-based SNPs

Table 14: Frequencies of commonly known alleles in African populations from this study and in Caucasians and Asians from literature sources

Table 15: Genetic diversity in CYP and NAT2 loci

Table 16: p-values for differentiation test of four ethnic groups based on HapMap SNP frequencies

Table 17: Pairwise comparison of populations (p-values based on allele frequencies from Table 14)

Table 18: AMOVA analysis on allele frequencies from Table 14

Table 19: Catalogue of SNPs analysed in this study

Table 20: Summary of catalogue

Table 21: Polymorphisms proposed for pharmacodiagnostic kit based on genes in this study

Table 22: Pharmacogenetics studies in African populations

LIST OF FIGURES

Figure 1: P4 medicine

Figure 2: Human CYPs

Figure 3: Sequence traces of novel non-synonymous SNPs

Figure 4: NAT2 haplotypes constructed from sequence and genotype data

Figure 5: Data analysis for population differentiation and relatednes

Figure 6: UPGMA tree calculated based on freqencies of 32 HapMap SNPs

Figure 7: PCA analysis using allele frequencies of DME genes in Africans, Caucasians and Asians

Figure 8: Un-rooted UPGMA tree showing population clusters based on frequencies of commonly known alleles

Figure 9: Resource building for pharmacogenetics in Africa.

Figure 10: Sample and genotype database

Figure 11: Sample and genotype database: Functions for basic statistical calculations

Figure 12: Complex pattern of haplotypes in CYP2C19, CYP2B6 and NAT2 gene regions

Figure 13: African map showing regions where pharmacogenetics studies (PG) have been carried out

LIST OF PUBLICATIONS

Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects.

Matimba,A., Del-Favero,J., Van Broeckhoven,C., Masimirembwa,C.

Hum Genomics. 2009 Jan 1;3(2):169-190. PMID: 19164093

Pharmacogenetics enables personalised therapy based on genetic profiling and is increasingly applied in drug discovery. Medicines are developed and used together with pharmacodiagnostic tools to achieve desired drug efficacy and safety margins. Genetic polymorphism of drug-metabolising enzymes such as cytochrome P450s (CYPs) and N-acetyltransferases (NATs) has been widely studied in Caucasian and Asian populations, yet studies on African variants have been less extensive. The aim of the present study was to search for novel variants of *CYP2C9* , *CYP2C19* , *CYP2D6* and *NAT2* genes in Africans, with a particular focus on their prevalence in different populations, their relevance to enzyme functionality and their potential for personalised therapy. Blood samples from various ethnic groups were obtained from the AIBST Biobank of African Populations. The nine exons and exon-intron junctions of the CYP genes and exon 2 of *NAT2* were analysed by direct DNA sequencing. Computational tools were used for the identification, haplotype analysis and prediction of functional effects of novel single nucleotide polymorphisms (SNPs). Novel SNPs were discovered in all four genes, grouped to existing haplotypes or assigned new allele names, if possible. The functional effects of non-synonymous SNPs were predicted and known African-specific variants were confirmed, but no significant differences were found in the frequencies of SNPs between African ethnicities. The low prevalence of our novel variants and most known functional alleles is consistent with the generally high level of diversity in gene loci of African populations. This indicates that profiles of rare variants reflecting interindividual variability might become the most relevant pharmacodiagnostic tools explaining Africans' diversity in drug response.

Establishment of a Biobank and Pharmacogenetics Database of African populations.

Matimba,A., Oluka,O., Ebeshi,B.U., Sayi,J., Bolaji,O.O., Guantai,A.N., Masimirembwa,C.M.

Eur J Hum Genet. 2008 Jul;16(7):780-3. PMID: 18382479 No abstract available

Objectives: The aim of this study was to establish a Biobank of the distinct ethnic groups in Africa and initiate a pharmacogenetics database of African populations. ***Methods:*** Ethical approval for the study was obtained from each of the countries from which samples were collected. Blood samples were collected from 50-100 adult volunteers from each of the major ethnic groups in Nigeria, Kenya, Tanzania, Zimbabwe and South Africa. Portions of each blood sample were used to prepare DNA, blot on filter paper or store at –80oC. Genotyping was initiated for alleles of major drug metabolising enzymes. A Database cataloguing the samples and the genotype results was designed using Microsoft Access and Visual Basics software packages. ***Results and conclusions:*** The biobank consist of one thousand five hundred samples from 9 ethnic groups (Yoruba, Hausa, Ibo, Luo, Kikuyu, Maasai, Shona, San, and Venda) in 5 African countries. The utility of the biobank was illustrated by studying the diversity of the African populations in drug metabolising genes thereby establishing a pharmacogenetics database of African populations.

Common FMO3 polymorphisms in 13 ethnic populations from Europe, East Asia and sub-Saharan Africa: frequency and linkage analysis.

Mao,M., **Matimba,A.**, Scordo,M.G., Günesa,A., Zengile,H., Yasui-Furukori,N., Masimirembwa,C., Dahl,M-L.

Accepted for publication: Pharmacogenomics. Sep 2009

Aims: To investigate intra- and inter-ethnic differences in three wide-spread (E158K, V257M, E308G) and two African specific (D132H, L360P) flavin-containing monooxygenase 3 (FMO3) polymorphisms. ***Materials and Methods:*** Allele frequencies were determined by TaqMan allelic discrimination assay in 2152 healthy volunteers from Europe (Swedes, Italians, Turks), East Asia (Japanese) and sub-Saharan Africa (9 ethnic groups covering Eastern, Southern and Western regions), followed by haplotype and linkage analysis. ***Results:*** Significant subpopulation differences ($P < 0.001$) in allele frequencies were found for E158K, V257M and E308G in Europeans and regional differences ($P < 0.01$) for D132H among Africans. No carrier of P360 was identified. G308 was always cis-linked to K158 with the compound variant (K158/G308) being found in high proportion (12.0% - 38.3%) of non-African subjects, but rarely (1.3%) among Africans. ***Conclusions:*** Distribution of functionally relevant FMO3 polymorphisms varies not only between ethnicities but also within. The K158/G308 variant may have potential clinical importance primarily in non-African populations due to its low prevalence in Africa.

1.INTRODUCTION

1.1. Human Genetics and Diversity

Medicine is changing faster than ever before. Its focus shifts from surprise to prediction, from healing to prevention, from one-size-fits-all to personalisation, from passive patients to participation. The goal is to maintain health rather than treat disease. Technology advances enable evermore early diagnosis and tailored therapy, empowering the human individual. In order to exploit this development in medical practice, pharmacogenetics is used to advise treatment, based on physical and digital resources such as biobanks and databases, warehousing and annotating information of populations (Figure 1).

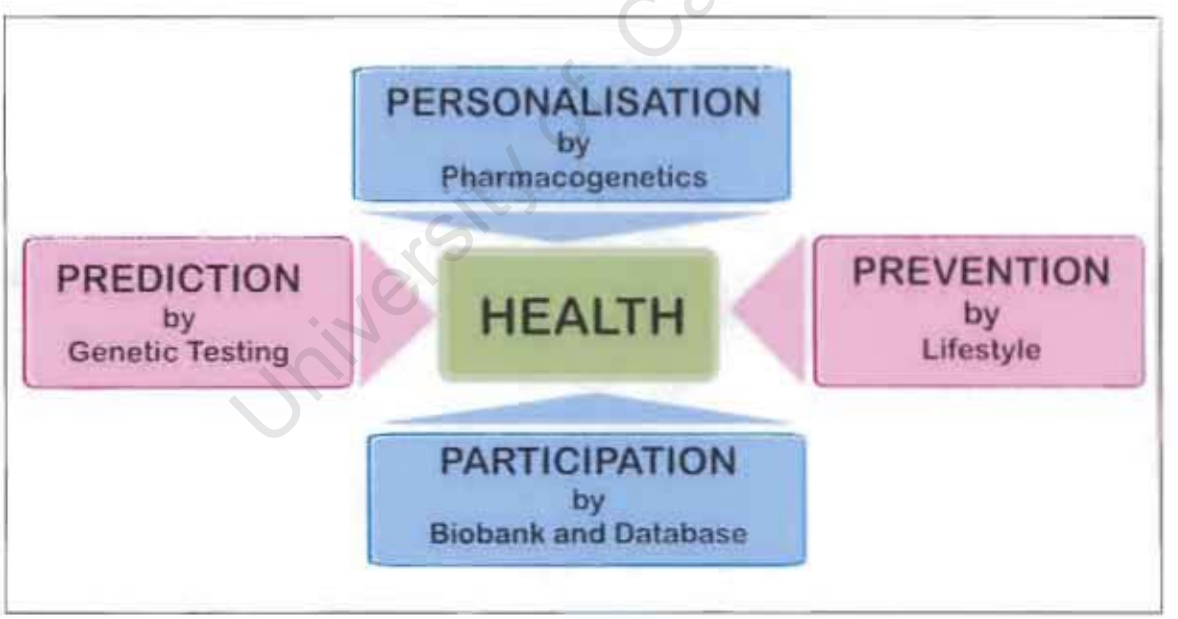


Figure 1: P4 medicine (term by Hood and Gallas, 2008: see 1.3.1)

While targeting populations for medical care, individual genetic profiles will enhance preventive measures and treatment outcome for patients. The elucidation of the human genome (Venter et al., 2001) and latest

advances in biotechnology (Mardis, 2008) indicate that sequencing of whole genomes combined with expression profiling will soon become available at reasonable cost (Hert et al., 2008). Consequently, healthcare will be transformed from empirical to evidence-based, informed by human genetics research.

1.1.1. Inheritance: genetics and evolution

A period of five years in the second half of the 19th century saw the description of the two main paradigms of inheritance: **genetics** and **evolution**. Their respective proponents, the Moravian/Bohemian monk Gregor Mendel (1822-1884) and the English naturalist Charles Darwin (1809-1882) knew nothing about each other, and their discoveries were quickly forgotten. Only by the start of the 20th century, when **chromosomes** were discovered and **genes** were described as the units of hereditary information, genetics was born as the science of heredity and biological inheritance (Bateson, 1905). At about the same time, the English physician Sir Archibald Edward Garrod (1857-1936) described inborn errors of metabolism and discovered inherited predisposition to drug response in humans (Garrod, 1909), yet the term **'pharmacogenetics'** was not used until the 1950s..

Classical genetics explains that offspring inherit genetic material from mother and father. Genes may occur in two or multiple, slightly different forms, or **alleles**, coding for the same trait. Each individual carries two alleles at each gene locus that are either the same (homozygous) or different (heterozygous). Genetic **polymorphism** describes the occurrence of different alleles in more than one percent of a population. If an allele occurs at less than one percent it is referred to as a 'rare allele'.

The genetic material, deoxyribonucleic acid (**DNA**), was first isolated in the 1940s (Avery et al., 1944). It contains the four nucleotide bases adenine (A), cytosine (C), guanine (G) and thymine (T); their order makes up the DNA sequence. The three-dimensional structure of DNA is a double helix of two strands of nucleotides held together by hydrogen bonds between the complimentary base pairs A-T and C-G (Watson and Crick, 1953).

Genes are molecularly defined as stretches of DNA that encode for proteins via messenger RNA (**mRNA**), an identical copy of the DNA sequence, except that thymine (T) is replaced by uracil (U). Protein expression starts with the production of mRNA (transcription) in the nucleus, its transport to the cytoplasm, mRNA processing and, eventually, its translation into protein. Every three nucleotides form a triplet or **codon**, the basis of the genetic code that determines the sequence of amino acids during the synthesis of proteins (Crick et al., 1961). This chain of events is called the 'central dogma of molecular biology' (Crick, 1970). Gene expression is under the control of regulatory elements and other genetic or environmental factors.

Structurally, a gene consists of a **promoter** region (for initiating gene expression), **exons** (amino acid coding sequences), **introns** (non-coding sequences) and untranslated upstream or downstream sequences or regions (**UTR**).

1.1.2. Phenotype and genotype

Whereas the observable physical or biochemical characteristics of an organism make up its **phenotype**, we refer to the DNA sequence as the **genotype**, the genetic blueprint for the expression of a phenotype. Variations or **mutations** within the DNA sequence alter

protein expression and/or activity. In spite of complex molecular processes that ensure the general integrity of the DNA sequence, mutations can be caused by physical damage to the DNA or by errors during its replication. While much of this variation goes unnoticed, mutations, together with physiological and environmental factors, can effect altered structures and levels of proteins, resulting in phenotypic differences. Some of these differences are of medical importance, causing resistance or predisposition to disease, the inability to digest certain types of foods and variation in drug response.

1.1.3. Variation in the human genome

Human genetic variation refers to differences in the occurrence of mutations between the genomes of individuals and populations. The DNA sequence of any two individuals differs once at approximately every 1,000 base pairs (Venter et al., 2001). The most common genetic variants are single base changes called **single nucleotide polymorphisms (SNP)** (Wang and Moult, 2001). SNPs in exons are called 'non-synonymous SNPs' if they cause amino acid changes or, together with base pair deletions or insertions, alterations of the reading frame. SNPs located in promoters, UTRs or splice-site junctions may affect gene expression.

Combinations of SNPs in a contiguous DNA segment are inherited together, forming **haplotypes**. Linkage disequilibrium (LD) means that SNPs or variants in close range have a higher chance of being found on the same haplotype. Patterns of linkage can be used for mapping disease, estimating diversity and evolutionary relatedness amongst individuals (Collins, 2009), and have been shown to differ across various populations (Conrad et al., 2006). The HapMap project aims to elucidate these patterns across major world populations such as

Caucasians, Asians and Africans, demonstrating genetic associations with common diseases (Manolio et al., 2008).

Structural variation in the human genome is very complex (Feuk et al., 2006; Marioni et al., 2008). Deletions and duplications lead to **copy number variants (CNV)** of genes or genome regions, resulting in their absence, reduced or amplified expression. Although human CNV polymorphisms are not yet fully ascertained, current efforts are directed toward a more comprehensive cataloguing and characterisation of these variants (Freeman et al., 2006; Redon et al., 2006; Shen et al., 2008). Integrated approaches have been used to determine SNPs in linkage disequilibrium with CNVs (McCarroll, 2008).

Other variations in the human genome include repetitive DNA sequences that occur as tandem repeat units (Scharf, 1995). Variable number tandem repeats are called **minisatellite** markers if they contain repeat units ranging from 8-50 base pairs. **Short tandem repeats (STR)**, or **microsatellites**, have a core repeat unit of 2-6 base pairs. Such variations may not confer a functional significance but may be in linkage with an SNP that causes an impact on gene expression and hence result in a disease condition. In such cases, association studies test whether an SNP or microsatellite is enriched in patients with disease compared with suitable control individuals. Candidate SNPs which alter the function or expression of the gene product can then be tested to see if they confer some phenotypic consequences. Microsatellites are also used in forensics for genetic fingerprinting of individuals (Hammond et al., 1994).

1.1.4. Genetic diversity of populations

Barriers such as major geographic features, language and ethnicity contribute significantly to genetic diversity of world populations (Belle and Barbujani, 2007). However, most of the genetic variation found in humans is shaped by the following processes (Templeton, 2006). First, random **mutations** are the primary source of genetic diversity in individuals. Second, **genetic drift** results in changes of allele frequencies by chance, 'fixing' certain alleles in a population and decreasing genetic diversity over time. Third, **gene flow** (or gene migration), as the migration of alleles between populations, increases genetic diversity within, while decreasing it between populations. Finally, **evolutionary forces**, when at work within and among populations, reflect the dynamics of genetic drift and gene flow between such populations. Because genetic drift is more obvious in smaller populations, the variance of neutral polymorphism frequencies increases among population groups. The polymorphisms that arose in one group are more likely to be restricted to this group. Thus the frequencies of genetic differences vary among populations of all kinds (Bamshad and Wooding, 2003).

These events have resulted in the diverse phenotypes observed across humans, such as skin colour, hair type and susceptibility to disease (Balaresque et al., 2007). For example, a mutation in the haemoglobin beta gene that causes red blood cells to become sickle-shaped, has apparently been enriched in certain populations as an adaptive response against the invasion by malaria parasites (Sabeti, 2008). This mutation is highly prevalent in west Africa, and homozygous individuals develop sickle cell anaemia. In contrast to such easily discernible phenotypes, more complex combinations of polymorphisms

(profiles) are used to infer population genetic diversity for use in medicine and anthropology (Harpending et al., 1998; Jorde and Olson, 2008). Technically, the assessment of genetic population diversity requires statistical analyses, the following selection of which was applied here (see 3.6.5).

The **Hardy–Weinberg (HW)** principle claims that genotype frequencies in a population remain constant, or are in equilibrium, from generation to generation unless specific disturbing influences are introduced (Hardy, 1908; Weinberg, 1908). HW assumptions are: there must be no mutation, no selection, no migration, a large population (no genetic drift) and random mating (no inbreeding). That means, HW equilibrium is impossible in nature, and deviation from it denotes the evolution of a species. Therefore, HW equilibrium is an ideal state that provides a baseline for measuring genetic change. Deviations from the HW equilibrium at a particular marker may indicate problems with the genotyping procedure or the population structure. To test for such deviations, Pearson's chi-square test (Pearson, 1900) is used for large populations, while **Fischer's exact test** (Fisher, 1922) is preferable if sample sizes are small (Raymond and Rousset, 1995; Wigginton et al., 2005).

F statistic is a measure of the difference between the mean heterozygosity among the subdivisions in a population and the potential frequency of heterozygotes if all members of the population mixed freely and non-assortatively (Hartl and Clark, 1997). In addition, an **isolation by distance** model assumes that genetic similarity between populations will decrease exponentially as the geographic distance between them increases, as geographic distance

limits the rates of gene flow (Relethford, 2004). Populations can be stratified by correlating genetic markers on different levels of diversity using **Principal Component Analysis (PCA)** (Novembre and Stephens, 2008; Paschou et al., 2007; Zhang et al., 2008). This allows the presentation of principal component maps overlaying the geographical distribution of markers (Handley et al., 2007; Prugnolle et al., 2005). Phylogenetic analysis with **Unweighted Pair Group Method Arithmetic (UPGMA)** can be used to infer evolutionary relatedness by way of calculating genetic distances based on allele frequency differences between population pairs and thus generating an evolutionary tree. Variance components are deduced in a hierarchical way by **Analysis of Molecular Variance (AMOVA)**, comparing population cluster, population, ethnic group and other variation levels to estimate the contribution of each level towards the overall diversity.

Classification of human populations along ethno-linguistic lines was first demonstrated by Cavalli-Sforza (Cavalli-Sforza et al., 2004), using HLA and blood group markers. Since then, frequency distribution of genomic markers such as microsatellites, SNPs and haplotypes has been applied to determine the genetic structure of the main world populations (Caucasians, Asians and Africans). Various autosomal markers were used to assess evolution of drug metabolism Sistonen et al., 2009.

Africa has been described as the cradle of mankind (Campbell and Tishkoff, 2008). Accordingly, high levels of genetic diversity were detected in Africans (Campbell and Tishkoff, 2008; Tishkoff and Verrelli, 2003). Using mainly microsatellite markers, ancestral African population clusters were recently refined, pinpointing the origin of

modern humans to south-western Africa (Tishkoff et al., 2009). Supported by data from mitochrondrial and Y chromosomal DNA analysis (Behar et al., 2008; Garrigan et al., 2007), population migration, long term ancestry and admixture have been estimated. Combining the genetic landscape, geography and prevalence of disease is expected to give guidance for estimating risk factors in populations of various geographical regions.

1.2. Africa

1.2.1. The African continent and its people

Africa covers 6% of the earth's total surface (20.4% of the land area) and is the world's second-largest and second most-populous continent, after Asia. There are 54 African countries, and a population size of approximately 934 million people accounts for 14.2% of the world's total population (United Nations World Population Division, 2008). The continent is surrounded by the Mediterranean Sea to the north, the Suez Canal and the Red Sea to the northeast, the Indian Ocean to the southeast, and the Atlantic Ocean to the west. Sub-Saharan Africa is a geographical term used to describe the area of the African continent south of the Sahara desert.

1.2.2. African languages and ethnicities

There are more than 2,000 languages in Africa, most of which belong to one of four language phyla: Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Khoisan (Gordon, 2005; Ruhlen, 1991). Based on the Ethnologue (www.ethnologue.com), these language classifications define ethno-linguistic clusters referring to ethnicities of populations as shown in Table 1.

Table 1: Some ethnic groups from eastern, western and southern Africa.

Region	Groups	Countries	Ethno-Linguistic Classification*	Est. pop. Size**
Eastern Africa	Kikuyu	Kenya	Niger Congo B-Bantoid	7 million
	Luo	Kenya, Tanzania, Sudan, Uganda	Nilo-Saharan-Nilotic	4 million
	Maasai	Kenya, Tanzania	Nilo-Saharan-Nilotic	< 1 million
	Sukuma	Tanzania	Niger Congo B-Bantoid	37 million
Southern Africa	Venda	South Africa, Zimbabwe	Niger Congo B-Bantoid	1 million
	Shona	Zimbabwe	Niger Congo B-Bantoid	10 million
	Zulu	South Africa	Niger Congo B-Bantoid-Nguni	10 million
	San	Botswana, Namibia, South Africa, Zimbabwe	Khoisan	<100 000
Western Africa	Hausa	Nigeria, Niger, Ghana	Afro-Asiatic-Chadic	35 million
	Igbo	Nigeria	Niger Congo A-Igboid	25 million
	Yoruba	Nigeria, Benin	Niger Congo A-Yoruboid	30 million

*Four ethno-linguistic classes are represented here
**Estimated population size according to CIA, 2009

The Niger-Congo language family is the largest in Africa (and probably in the world) in terms of its number of languages. Ethnic groups from this family can be split into two sub-families. The Yoruba and Igbo belong to sub-family A, which is particularly populous in west Africa, with population sizes of 30 and 25 million, respectively, in Nigeria.

However, the major branch of the Niger-Congo family are the Bantu (subfamily B), which consist of over 400 ethnic groups, supposedly originated from central Africa, spreading across east and southern Africa, and are estimated at a population size of 240 million (http://en.wikipedia.org/wiki/Bantu_peoples). In most eastern and southern African countries, Bantu ethnicities are the most populous, making up 70% of the Kenyan population (of 39 million), of which the

Kikuyu are the largest group (22%; (CIA, 2009). Ethnic groups such as Sukuma make up 90% of the Tanzanian population (of 41 million). In southern Africa, Bantu ethnicities include Shona, Zulu, Xhosa, Venda and Ndebele, with varying population sizes. While Shona are the majority in Zimbabwe (82% of 11 million), Venda make up only 2.5% of the South African population (49 million).

The Hausa (Afro-Asiatic) form one of the largest ethnic groups of west Africa, being prevalent in Nigeria, Niger, Benin and Cameroon, amongst other places. The estimated population size of Hausa is 30 million in Nigeria alone (CIA, 2009).

The Nilo-Saharan language family includes Nilotics and Cushite subfamilies, which mostly populate central-east-north and eastern parts of Africa. Nilotics include Luo, Kalenjin, Maasai in Kenya and are also found in Tanzania, Uganda and Rwanda. The Dinka of Sudan represent Nilotic people in north-east Africa.

1.2.3. Disease burden in Africa

The African continent is a diverse landscape of human genetics as well as disease. Battling with the suffering from HIV/AIDS of over 22 million people in sub-Saharan Africa alone (WHO, 2008), and several anti-retroviral drugs on the market, the search for an HIV vaccine continues (http://chi.ucsf.edu/vaccine/vaccines?page=vc-00-00). Currently, anti-retroviral treatment is often based on a combination of three drugs (e.g. lamivudine, zidovudine, abacavir). Their efficacy and **adverse drug reactions (ADR)** are monitored by CD4+ counts and determination of viral load and drug plasma concentrations. ADRs often result in reduced compliance of patients, leading to survival of drug-resistant strains (Ivanovic et al., 2008; Laurent et al., 2008) and

recommendation of alternative treatment (Hammer et al., 2008). However, safe and effective medicines, as well as monitoring services, are often expensive and unavailable in parts of Africa, complicating the fight against this disease.

With about 9 million new cases and nearly 2 million deaths every year, tuberculosis is the number one killer of people co-infected with HIV (WHO, 2005). Timely diagnostic methods and follow up of treatment present major challenges for resource-poor nations. The situation is compounded in cases of **multi-drug resistant (MDR)** and extensively drug-resistant (XDR) strains, for which the two most powerful first-line drugs, isoniazid and rifampicin, are rendered ineffective (Banerjee et al., 2008; Kim et al., 2007). Alternative treatment for drug-resistant strains usually means less-potent drugs, which may also trigger severe ADRs. *Mycobacterium* pharmacogenetics maybe used to determine the markers conferring resistance to this micro-organism (Warren et al., 2009).

In 2004, more than 1 million malaria-caused deaths were reported in Africa (WHO, 2004). Effective treatment has been established with artemisinin-amodiaquine-based combination therapies. These drugs are more expensive but also more effective than alternatives such as chloroquine, whose efficacy is reduced due to drug resistance (Sirima et al., 2009; Yeung et al., 2004). Recently, emerging artemisinin-resistant strains have been found in Asia, hindering progress in controlling the parasite globally (Epstein and Thenabadu, 2009). The high prevalence of the above diseases has a tremendous impact on social and economic development of African nations, as therapeutic interventions are still inaccessible in many African regions.

Cancer diagnosis is on the increase in Africa, with challenges in access to adequate treatment. For example, it was reported that Zimbabwean cancer patients experience low survival rates due to economic constraints (Gondos et al., 2004). In addition, increase in infectious diseases results in increased cases of conditions such as Karposi sarcoma and cervical cancer. Epidemiological studies have revealed the high prevalence of breast, prostate and colorectal cancers (McLary, 2009).

Due to changes in diet and environment, 'lifestyle' disorders such as obesity, cardiovascular diseases and type II diabetes are on the rise in Africa (Frost and Sullivan, 2009; Steyn et al., 1992).

Although some treatment is available for most conditions, its efficacy and safety is often a challenge. Limited, mostly government-directed funds, hard choices on imports, and counterfeit medicines are major problems. In order to improve on this unfortunate state of affairs, advances in disease diagnostics and surveillance are paramount for advising appropriate treatment. Using pharmacogenetic strategies, drug response can be gauged and tailored to the needs of African populations, attempting individualised therapy.

1.3. Personalised Medicine

Traditional clinical approaches would combine an individual's personal medical record and family history, together with data from imaging and laboratory tests to diagnose and treat disease. With the complete sequence of the human genome (Venter et al., 2001) another parameter was added, which has the potential of changing the fundamentals of medicine. Benefiting from advanced molecular technology and new information resources, the experimental discipline of **genomics** promises matching genotype and phenotype, enabling personalised medicine (Laberge and Burke, 2008).

1.3.1. Genomics-informed medicine

Genomics employs the analysis of an individual's genome by searching for variations which make the individual unique in the way they are susceptible to disease or respond to medicine (Guttmacher and Collins, 2002). Advances in translational genomic research have the following specific aims in medicine: to identify and monitor individuals at high risk from disease, to design most effective drugs, to prescribe the best treatment for each patient, and to avoid ADRs.

Recently, the term **'P4 medicine'** was coined, describing healthcare of the future as **predictive**, **preventive**, **personalised**, **participatory** (Figure 1, Hood and Gallas, 2008). Personalised medicine refers to individual-based therapy, offering the right treatment to the right person at the right dose. Based on the fact that 'blockbuster' drugs are only effective in about 40% of people (PWC, 2005), increasing cost and risk of drug development as well as public demand for more effective and safer drugs, even the pharmaceutical industry has now embarked on a personalised medicine campaign.

The central concept is to focus treatment on patient groups that actually respond to it by combining every therapy with a **companion diagnostic (Dx)**, so that only positive-testing patients would receive the treatment (Allison, 2008). Technically, this is based on molecular screening for variants that result in a certain phenotype. Linking a specific diagnostic with a specific regimen of prescribed drugs, e.g. HercepTest/trastuzumab (Dako, Carpinteria, CA) for Her-2 positive breast cancer patients, ensures the development of drugs highly targeted to patients who are most likely to benefit, thus optimising treatment outcome. Anti-cancer drug Herceptin is specifically tailored for tumours over-expressing human epidermal growth factor receptor-2 protein (Her-2/neu) (Vogel et al., 2002). Therefore, expression profiles of cancer patients are assessed to identify individuals who are most likely to benefit from the drug (van't Veer and Bernards, 2008).

This type of diagnostics is collectively referred to as **molecular diagnostics**, the technical foundation of personalised medicine. Its major clinical applications are outlined under the following three topics (1.3.1.1-1.3.1.3), which are based on **genetic testing** (1.3.1.4).

1.3.1.1. Monogenic and complex disorders

Monogenic disorders (also known as Mendelian disorders) such as cystic fibrosis, sickle cell anemia, thalassemia and many others, are determined by mutations in a single gene. On the other hand, complex disorders may involve several loci (Peltonen and McKusick, 2001) and biochemical pathways (Dempfle et al., 2008) as well as environmental interactions. Complex disorders include common diseases such as obesity, diabetes, hypertension and cardiovascular diseases, cancer

and neurological disorders such as Alzheimer's, Parkinson's and Huntington's disease (Belmont and Leal, 2005; Motulsky, 2006).

The Online Mendelian Inheritance in Man (OMIM) is a regularly updated catalogue of human genes and genetic disorders, based on (McKusick, 1998). It provides an essential resource of genetic markers for disease risk prediction and preventive advice, enabling pre-symptomatic identification of disease by genetic testing.

1.3.1.2. Genetic susceptibility to infection

It has been shown that susceptibility to infection as well as progression to disease can be genetically determined (Anastassopoulou and Kostrikis, 2003; Casanova and Abel, 2007). For example, individuals homozygous for the chemokine receptor CCR5 delta32 allele show some protection against HIV infection, or express less severe disease if infection occurs (Carrington et al., 1999). Genetic susceptibility to tuberculosis in Africans appears to be polygenic (Fitness et al., 2004), associated with a locus on 8q12–q13 (Baghdadi et al., 2006), amongst others.

1.3.1.3. Drug response as a complex trait

Pharmaceutical drug response is considered a complex trait (Goldstein, 2005), involving the interplay of genetic and physiological factors. The elucidation of this complexity proves a formidable challenge, similar to common diseases (Vella and Camilleri, 2008). Progress in this field is driven by advances in pharmacogenetics, as outlined below (see 1.4.) and used in this study.

1.3.1.4. Genetic testing and prediction

In order to diagnose genetically determined conditions, DNA sequence information needs to be obtained from human subjects. Genetic testing laboratories are established in most countries, with major public (Mayo Clinic, Rochester, MN; GENDIA, Antwerp, Belgium) and private organisations (Genzyme, Cambridge, MA; Clinical Data, Newton, MA; LabCorp, Burlington, NC; Bioscientia, Ingelheim, Germany) operating globally.

Referring to the full potential of personalised medicine, **Direct-To-Consumer (DTC)** personal genetic testing is now offered to the general public by some companies (deCODE genetics, Reykjavik, Iceland; 23andMe, Mountain View, CA; Navigenics, Foster City, CA; Knome, Cambridge, MA), providing whole genome profiles for genetic risk assessment.

1.3.2. Resources for genomic research

Genetic testing entails technology advances that enable the detailed analysis of individual DNA sequences, collectively called **genotyping**. Ranging from single-gene analysis to whole genome sequencing, genotyping and re-sequencing are used to analyse variation in individuals and at population level.

The central goal of genomic research is the discovery of **diagnostic markers** for predicting, preventing, treating and monitoring disease and susceptibility to infections. This requires the annotation of genetic variants in a population with clinical and lifestyle data, interactions with environmental and other risk factors, an endeavor collectively called **phenotyping**. As these efforts need to be undertaken on an individual basis, a **sample** of each member of the population needs to

be taken, processed and stored in a physical collection, called a **biobank**, linked to annotating information in an electronic **database**.

1.3.2.1. Methods in genomics

Whereas genotyping is used to test for known variants such as SNPs, single base insertions and deletions as well as CNVs, **re-sequencing** means scanning target regions for variations compared to a reference sequence. New mutations discovered by re-sequencing can later be used for genotyping to determine the alleles carried by an individual.

Technically, Sanger-based (Sanger et al., 1977) capillary sequencing (Applied Biosystems, Foster City, CA) remains the most versatile re-sequencing method, featuring long reads and high levels of accuracy. The main advantage of new generation sequencing technologies (based on hybridisation, microarrays, single-molecule and/or nanopore technologies) is their higher throughput (Shendure and Ji, 2008). Currently, there is a race to develop genome sequencing methods that will be ever faster and more affordable. Several companies such as Roche Applied Science (Indianapolis, IN), Illumina (San Diego, CA), Applied Biosystems (Foster City, CA), Helicos (Cambridge, MA), Pacific Biosciences (Menlo Park, CA), Complete Genomics (Mountain View, CA) and Oxford Nanopore (Oxford, UK) have made great progress, and it might well be a realistic outlook to sequence a whole human genome for less than USD 10,000 by the end of 2009.

In recent years, a great deal of effort has been devoted to developing accurate, rapid, and cost-effective technologies for genotyping, yielding a large number of distinct approaches (Kim and Masra, 2007). Methods for assaying DNA variation require two important steps: (i) discriminating the variation and (ii) detecting the signal. **Restriction**

Fragment Length Polymorphism (RFLP) is a difference in homologous DNA sequences that can be detected by the presence of fragments of different lengths after digestion of a DNA sample with specific restriction endonucleases. Genotyping using this method involves **polymerase chain reaction (PCR)** to amplify the DNA target sequence, followed by RFLP analysis. Detection of fragments involves their separation on an agarose gel and staining by fluorescent dyes. This is the most traditional method for detecting polymorphisms. Although high levels of accuracy are possible, this method suffers limitations in terms of throughput.

Large scale genotyping can be achieved using high-throughput methods such as **TaqMan** (Shen et al., 2006). Developed by Applied Biosystems (Foster City, CA), Taqman genotyping involves quenched fluorescent probes, which are designed such that they anneal within a DNA region amplified by a specific set of primers. During PCR of this region, degradation of the probe releases the fluorophore from it, breaking the close proximity to the quencher and thus allowing fluorescence, which is detected in the thermal cycler and is directly proportional to the amount of amplified DNA template.

Sequenom's (San Diego, CA) primer extension technique is a two step process that first involves the hybridisation of a probe to the bases immediately upstream of the SNP nucleotide, followed by a 'mini-sequencing' reaction, in which DNA polymerase extends the hybridised primer by adding a fluorescently labelled base that is complementary to the SNP nucleotide. Detection is by matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS). Because multiple polymorphisms can be detected in a single

reaction, the technique provides a cost-effective and efficient method for high-throughput genotyping (Bray et al., 2001).

Taking throughput another level up, DNA microarrays allow global genotype analyses in chip format. This has enabled **genome wide association studies (GWAS)**, assessing genotypes or gene expression linked to diseases (Hardy and Singleton, 2009) such as type II diabetes (Rotimi et al., 2004) and cancer (Tomlinson et al., 2008). However, limited knowledge on common disease phenotypes and their interaction with genotypes keeps predictive rates still as low as 1% (Feero et al., 2008; Kraft and Hunter, 2009).

In pharmacogenomics, expectations are somewhat more optimistic (Roses et al., 2007). GWAS is expected to deliver on the goal of personalised drug therapy in the near future, as exemplified by studies on anti coagulant therapy with warfarin (Cooper et al., 2008) and other ongoing trials on anti-hypertensive therapy with beta-blockers (O'Shaughnessy, 2009). Although GWAS is claiming much attention for unveiling ADR markers (Crowley et al., 2009; Guessous et al., 2009), it has been argued that the polymorphisms analysed are less representative of certain ethnic groups such as Africans. This is based on the Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA) 1 million SNP chips, which may not take into account the contribution of rare functional SNPs (Gurwitz and McLeod, 2009). Therapeutic response is considered as complex trait and hence GWAS application in pharmacogenetics requires large patient and control samples (Crowley 2009).

High-powered computing resources and informatics tools have been developed to handle the extraordinary large amounts of data generated in human genomics (Joshua and Boutros, 2008; Karchin, 2009; Teufel et al., 2006).

As genomic analysis is getting more and more straightforward, the correlation of genome data with phenotypic/clinical information is of increasing importance (Snyder et al., 2009). Therefore, the collection of clinical samples and data in biobanks and genetic databases is vital to realise pharmacogenetics' potential for personalised medicine.

1.3.2.2. Biobanks

A modern biobank is the interface between tissue donors—either patients or healthy individuals—in the clinical care setting and scientists performing biomedical research in an academic or pharmaceutical industry setting. The biobank's role is to act as a secure yet effective bridge for samples and data to move between the two environments (www.biobankcentral.org).

Also known as **biorepositories**, biobanks house biospecimens, which can be tissue, cells, blood, serum, urine, buccal swabs or other body fluids. The aim is to preserve the biological material as scientific resource and reference. Modern biobanks are enterprise-like facilities, including staff and management, ethical and legal oversight, financial systems, storage facilities, laboratories, security and computer information systems to fully implement their operations. Commonly, a biobank features high-throughput capabilities, such as robotics and automated micro-quantity liquid handling, to isolate and handle biospecimens and their chemical components such as DNA.

Laboratory Information Management Systems (LIMS) have become standard practice ensuring seamless links between samples, experiments and data. The Karolinska Institute Biobank, in partnership with IBM, have introduced a modified core system known as **Biological Information Management System (BIMS)**. Both systems involve the maintaining of records of sample collection, transportation, storage, analytical methods used, results and other downstream activities.

Today, such bioresources exist in a variety of settings, such as academic and medical institutions, pharmaceutical and biotechnology companies. They can also be stand-alone organisations, including independent companies (both for-profit and non-profit) that can provide biobanking as an outsourced service or can serve as a broker of biological materials to other researchers. Old collections, whereby a researcher would have collected blood samples from individuals or families for diagnostic purposes over time, may be used for 'retrospective' studies. In contrast, 'prospective' biobanks assess and take samples from participants at the start of a study and then follow their medical or lifestyle record over subsequent years, even decades. Currently, both prospective and retrospective biobanks have become a major source of information for studies linking genetics with disease networks, lifestyle and environment. Administered at national, regional and international levels, a variety of population biobanks have been successfully launched around the world (Table 2).

Biobanking is an important tool for translational research aimed at new biomedical and clinical practices, diagnostic techniques and preventive treatments. Polymorphisms predisposing individuals to diseases such

as breast cancer have been discovered by analysing population samples in biobank consortia (Antoniou et al., 2008; Chenevix-Trench et al., 2007). Using the Utah Population Database, the most popular database tracking family history, medical records and cause of mortality, it is possible to follow pedigrees spanning several generations. Such resources have enabled determination of disease loci from multi-genealogical family data (www.hci.utah.edu/).

Table 2: Examples of Biorepositories

Name (Country/region)	No. of samples
National Programmes	
Biobank Japan	500,000
BioHealth (Norway)	500,000
Cartagene (Canada)	20,000
China Biobank Project	5 million
deCODE (Icelandic Biobank)	270,000
Estonia Genome Project	1 million
LifeGene (Sweden)	500,000
National Bank of Finland	200,000
UK Biobank	500,000
Utah Population Database	6.5 million
Disease	
African American Biobank (USA)	25,000
European Prospective Investigation into Cancer and Nutrition (EPIC)	24,000
Gambian Biobank	>57,000
Fertility	
Women's Biobank of India	in planning phase
Newborn/Families/Registries	
Danish Newborn Screening	1.8 million
GenomEUtwin (Finland)	520,000
Joondalup Family Health Study (Australia)	80,000
Networks	
Asian Cohort Consortium	
Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)	
Population Project in Genomics (P3G)	

The use of biobanks ranges from drug discovery to biomarker development to clinical trials, to stratify patients based on their genetic characteristics, disease markers and likelihood of response to the tested substance. Together with the pharmaceutical company Roche

(Basel, Switzerland), deCODE genetics (Reykjavik, Iceland) have used their biobank and population database for the discovery of drug targets of a wide range of diseases such as myocardial infarction, stroke, thrombosis and embolism, asthma, hypertension, schizophrenia, osteoporosis, obesity, type II diabetes, prostate cancer and osteoarthritis (www.decode.com/ClinicalDevelopment.php). Other pharmaceutical companies such as GSK (Brenton, UK) are working with the Joondalup Family Health Study in Western Australia (www.jfhs.org.au) to enhance their drug development programmes. The commercial biobanks of GeneLogic (www.genelogic.com) and BioRep (www.biorep.com) offer collection and storage of samples from individuals participating in clinical trials to pharmaceutical companies.

The importance of biobanks is further highlighted in situations where drug surveys are helped by retrospective analysis of samples in various populations. For example, AstraZeneca's (London, UK) anti-coagulant drug Exanta (Ximelagatran) was removed from the market due to adverse effects on hepatic function. Based on biological samples from patients taking the drug (Kindmark et al., 2007; Wilkinson, 2009), a retrospective pharmacogenetic case control study allowed to determine the genetic cause of those reactions.

1.3.2.3. Genetic databases

A major advantage of biobanks is their convenient and cost-effective storage of information as DNA. In contrast, whenever this information is being extracted, it needs to be stored electronically. Therefore, the huge amounts of data that are generated by genome projects require extensive information technology infrastructure to set up specialised data repositories (databases) and bioinformatics tools for analysis and interpretation. The two main global hubs of this effort are the National

Center for Biotechnology Information (NCBI), Bethesda, MD, and the European Bioinformatics Institute (EBI), Hinxton, UK, hosting the data repositories GenBank (www.ncbi.nlm.nih.gov/Genbank) and EMBL Nucleotide Sequence Database (www.ebi.ac.uk/embl), respectively. In addition, there are local data centres in research, healthcare and industry wherever DNA sequence information is used.

In conclusion, these investments are bound to grow exponentially with progress in genomic technology. However, similar to the development of the personal computer in the 1980s, fast and affordable DNA sequencing will help decentralise data resources, putting genome information in patients' and consumers' hands. In fact, this will further increase the importance of biobanks too, enabling to extract sequence information from stored samples quickly.

In summary, genomics is driving the advancement of personalised medicine, yet it is still too early for its adequate translation into routine medical practice (Scheuner et al., 2008). Challenges need to be addressed, ranging from validated genotype-phenotype associations to education of patients and clinicians (Lesko, 2007). However, the new paradigm of genomics-informed medicine is steadily taking hold across the board of clinical specialities. This development is raising awareness that a long established discipline of medical inquiry is actually becoming personalised medicine's most advanced field, encompassing all others: pharmacogenetics.

1.4. Pharmacogenetics

1.4.1. Drug response and pharmacogenetics

Pharmacogenetics is the study of the genetic factors that influence an individual's reaction to a drug. Pharmaceutical drugs and chemicals are foreign compounds, or **xenobiotics**, to which humans are exposed therapeutically, occupationally or through the diet. Xenobiotics have to undergo various transformations to effect treatment and facilitate their removal from the body. The absorption, biotransformation and elimination of xenobiotics by enzymes is important for safeguarding the body against toxic compounds and reactive species, which may modify cellular constituents and result in toxic reactions.

Variation exists in the levels and activities of biotransformation enzymes and carrier proteins due to both genetic and non-genetic factors such as environmental and physiological conditions. Polymorphic drug receptors may have reduced affinity. For pharmaceuticals, the effects of genetic polymorphisms range from drug inefficacy, excessive concentration of reactive intermediates or products to exaggerated immunological response or hypersensitivity reactions. Pharmacogenetics aims to identify and categorise the genetic factors that underlie these differences and apply the results in clinical practice.

1.4.1.1. History and current advances in pharmacogenetics

The English physician Sir Archibald Edward Garrod is credited with the first pharmacogenetic discoveries. He found that people with urinary disorders varied in drug response and later described 'taste blindness' in individuals who could not taste phenylcarbamide (PTC). Such

variation was later linked to people's ethnic backgrounds (Fox, 1932; Snyder, 2009).

The term 'pharmacogenetics' was introduced in the 1950s (Vogel, 1959) after glucose-6-phosphate dehydrogenase deficiency was discovered in African American soldiers suffering from haemolysis after taking the anti-malarial drug primaquine (Carson et al., 1956). Other early drug response episodes were described with succinylcholine, an adjunct to anaesthesia (Lehmann and Ryan, 1956), and isoniazid (Mitchel and Bell, 1957; Evans et al., 1960), which was and still is one of the most effective anti-tuberculosis drugs. Such reports increased in number during the 1970s with molecular characterisation studies being more prominent in the 1980s (Meyer, 2004). Whereas pharmacogenetics studies originally focused on monogenic traits, the discovery of entire pathways and disease mechanisms is increasingly becoming an important research area (Weinshilboum and Wang, 2006).

1.4.1.2. Properties of drug response

Pharmacodynamics (PD) describes the biochemical and physiological effects of drugs either on the body directly or on microorganisms or parasites within or on it, the mechanisms of drug action and the relationships between drug concentration and effects (Lees et al., 2004). This enables the formulation of dose-response relationships of drug potency and efficacy. The **therapeutic index** describes the difference between the dose required to effect treatment and that producing unwanted and possibly toxic effects.

Whilst pharmacodynamics explores what a drug does to the body, **pharmacokinetics (PK)** investigates what the body does to the drug.

Tied together, PK-PD describe the role of proteins either as pharmacokinetics factors, determining the concentration of a drug that reaches its target, and pharmacodynamic factors, influencing the drug target itself. When a drug is administered to the body it undergoes **absorption** into the blood stream, **distribution** to its site of action by protein carriers, **metabolism** by enzymes and **excretion**. Taken together and abbreviated as **ADME**, these processes result in individual variability of drug response. In other words, two individuals who take the same drug will not have the same effective plasma concentration. In addition, drugs can cause adverse chemical reactions, summarised as **toxicity**, completing the acronym as **ADMET**. Other factors that can exaggerate pharmacokinetic variability include drug-food interactions, drug-drug interactions, drug-disease state, gender, pregnancy and genetic differences. These factors must be assessed in detail before treatment, in order to ensure that individuals receive the right dosage to achieve potency whilst minimising adverse drug reactions (ADRs).

1.4.2. ADMET genes

Based on the variation in individuals' response to therapy, a person's genetic makeup or genome can be used to predict the probability of a certain drug response. Whereas pharmacogenetics starts out from an unexpected drug response, looking for a genetic cause, **pharmacogenomics** refers to the study of all known genes and their products that determine drug behavior. Therefore, today's high-paced advancement of genotyping technology feeds directly into the progress of pharmacogenomics, aimed at individual whole genome information for personalising treatment. Although both terms have been used interchangeably, it seems that pharmacogenetics is more commonly referred to as the study of single genes and their effects on inter-

individual differences in drug metabolising enzymes (DMEs) or receptors, while pharmacogenomics represents a whole-genome view of drug response variability.

Genes encoding ADMET proteins are highly polymorphic, hence of great interest to pharmaceutical discovery and development for maximising drug efficacy, minimising toxicity and selecting responsive patients for clinical trials. Besides DMEs, ADMET proteins include receptors and drug transporters such as the multi-drug transporter 1 (*ABCB1*). This protein belongs to a family of ATP-Binding Cassettes (ABC) and is involved in the transport of most drugs across the intestine into the blood stream. Found to be highly polymorphic, *ABCB1* has been associated with variation in the distribution of substrates including anti-cancer agents, cardiac drugs and HIV protease inhibitors (Fung and Gottesman, 2009; Kimchi-Sarfaty et al., 2007).

During the early years of pharmacogenetics, studies focused on outlier samples with the hope of identifying inherited variation in one or a few enzymes involved in the metabolism of a particular drug. With the advance of genomic high-throughput technology, this is now possible in a large-scale, prospective way by screening volunteer populations for polymorphisms, after which candidate variants can be tested in patients or individuals with an abnormal phenotype. In effect, the single gene/SNP approach has progressed to a situation where many genes in pathways can be analysed simultaneously towards determining underlying molecular mechanisms responsible for pharmacodynamic and pharmacokinetic variation.

1.4.3. Variation in drug metabolism

Drug metabolism includes activating/deactivating a drug, attenuating its biological activity and accelerating its clearance from the body. Drug metabolising enzymes act as shields against exposure to toxic chemicals or xenobiotics and can be divided into two major categories (Evans and Relling, 1999). The oxidative or phase I drug metabolising enzymes catalyse the introduction of an oxygen atom into substrate molecules, resulting in hydroxylation or demethylation. Examples of phase I enzymes are the cytochrome P450s (CYP), epoxide hydrolases (EH) and flavin mono-oxygenases (FMO). Five of the human CYPs (1A2, 2C9, 2C19, 2D6, 3A4) are involved in 95% of the CYP-mediated metabolism of drugs (about 75% of drug metabolism) (Rendic, 2002; Guengerich, 2003). The conjugative or phase II enzymes catalyse the coupling of endogenous small molecules to substrates to make them more soluble and hence easily excretable. Phase II enzymes include thiopurine S-methyltransferases (TPMT), uridine glucuronyl-transferases (UGT), N-acetyltransferases (NAT), glutathione S-transferases (GST) and sulphonyl tansferases (SULT).

The genetic polymorphism of **drug metabolising enzymes (DME)** is one of the most important and best understood causes of inter-individual variability in drug response. Variants identified in the coding genes have been associated with abolished, reduced, and sometimes increased enzyme activity. Effectively, this divides the population into **slow metabolisers** and **rapid metabolisers**. Given that the nature of genetic variation can affect enzyme activity at various levels, the phenotypic spectrum can be divided further into ultrarapid, rapid, intermediate and slow metabolisers. Ultrarapid and rapid metabolisers are likely to require higher doses of a drug due to their ability to

eliminate the drug from the body faster, resulting in sub-therapeutic levels at the target site. This is the case if the drug is administered in therapeutically active form. Slow metabolisers may need to have their dosage decreased to avoid accumulation of the drug, resulting in toxicity. This may be the case for intermediate metabolisers too, particularly when exposed to drugs with a narrow therapeutic index.

1.4.4. Drug metabolising enzymes

1.4.4.1. Cytochrome P450 (CYP)

The cytochrome P450s are named after the spectral absorbance peak of their carbon monoxide-bound species at 450 nm. These enzymes are referred to as mixed-function oxygenases, transforming lipophilic substrates into more water soluble compounds by using NADPH as a coenzyme and oxygen as a substrate (Garret and Grisham, 1995). In addition, CYPs catalyse the oxidation of non-activated C-H bonds (Isin and Guengerich, 2008). Generally, CYP enzymes are remarkable both for the diversity of reactions they catalyse and the range of chemically-dissimilar substrates upon which they act (Danielson, 2002).

CYPs consist of 18 families sharing >40% amino acid identity and 43 subfamilies of >55% internal amino acid identity. The human genome project has identified at least 62 CYP genes (57 intact and 5 pseudogenes). CYP genes are commonly abbreviated such as '*CYP2D6*' or '*CYP3A4*', of which the first number is the family and the letter is the subfamily, followed by a number that represents the isoform in that particular subfamily (Figure 2).

Located in the endoplasmic reticulum of cells, CYPs are highly concentrated in the liver, where major detoxification processes take place (Hayes, 1999). CYP expression is also observed in other tissues such as brain, kidneys and intestines. In the liver, CYP3A4 is the most abundant (~30%) followed by CYP1A2 and CYP2C (15-20% each), while the rest of the CYPs such as CYP2B6, CYP2D6 and CYP2E1 make up less than 5% each (Shimada et al., 1994). As CYPs are pivotal in drug metabolism and toxicity, they represent a major concern in drug development (Smith et al., 1996); their substrates have been categorised (www.medicine.iupui.edu/flockhart/table.htm).

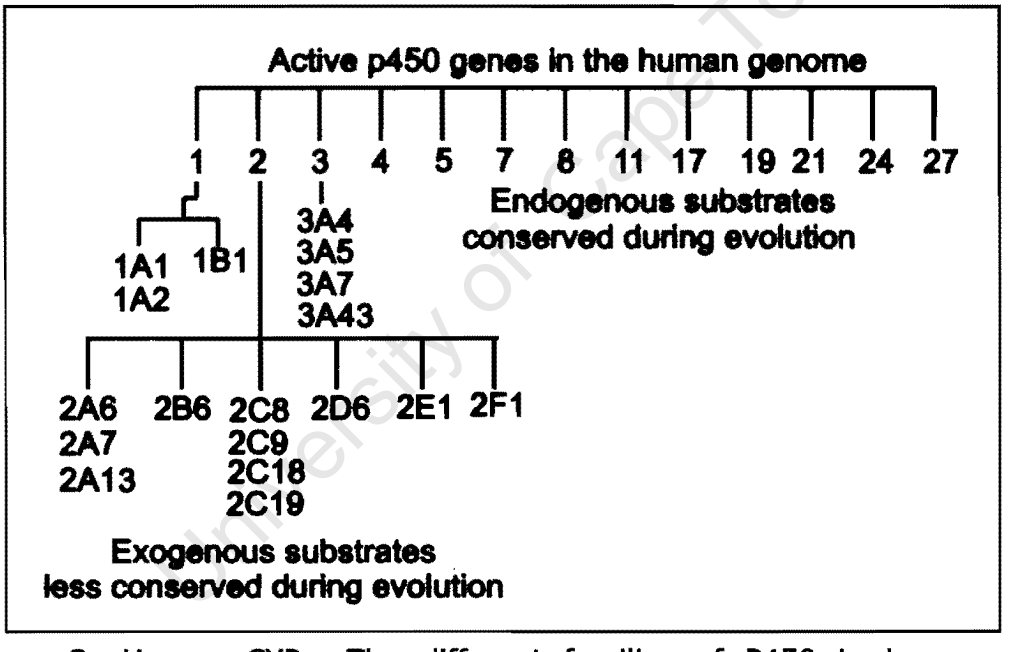


Figure 2: Human CYPs. The different families of P450 in humans are represented by vertical lines, those of families 1–3 are enzymes involved in xenobiotic metabolism, while those in higher families metabolise endogenous substrates. The subfamilies are represented by a letter, while isoforms are given Arabic numbers. (Ingelman-Sundberg and Rodriguez-Antona, 2005)

CYP3A4 contributes most to the metabolism of clinical drugs (~40%). Other major contributions come from *CYP2C9* (17%), *CYP2D6* (15%), *CYP2C19* (10%), *CYP1A2* (9%), *CYP2C8* (6%), and *CYP2B6* (4%) (Zanger et al., 2008).

The highly polymorphic nature of CYPs causes inter-ethnic and inter-individual differences. Alleles of CYP genes are assigned names such as *CYP2D6*17*, in which *17 refers to a polymorphism within the sequence of that particular gene. An allele nomenclature committee keeps track of these names and new allele assignments (www.cypalleles.ki.se).

1.4.4.1.1 CYP2B

This sub-family has only one member in humans, *CYP2B6*, one of the less well studied human CYPs. *CYP2B6* is primarily expressed in the liver, involved in the first-pass metabolism of ingested drugs, but has also been detected in several extra-hepatic tissues including brain, kidney and intestine. Human *CYP2B6* is estimated to account for a minor portion (<1%) of total hepatic CYP content and to have a minor function in human drug metabolism. Recent studies, however, indicate that the average *CYP2B6* content in human liver is approximately 3–6% of the CYP pool and that *CYP2B6* plays a critical role in the biotransformation of several clinically important drugs (Wang and Tompkins, 2008; Zanger et al., 2007). These substrates include cytostatics (cyclophosphamide), anti-HIV drugs (efavirenz and nevirapine), anti-depressants (bupropion), anti-malarials (artemisinin), anesthetics (propofol) and synthetic opioids (methadone).

The *CYP2B6* gene, together with the expressed pseudogene *CYP2B7*, is located within the 350kb CYP2ABFGST gene cluster on chromosome 19 that contains genes and pseudogenes of the CYP2A, 2B, 2F, 2G, 2S and 2T subfamilies. *CYP2B6* is highly polymorphic with 28 characterised alleles, over 50 haplotypes and more than 100 described SNPs. Contrary to the conventional path to polymorphisms such as in

CYP2C19 and *CYP2D6*, which were discovered via adverse drug reactions, pharmacogenetic analysis of *CYP2B6* was initiated via reverse genetics approaches, followed by functional and clinical studies (Zanger et al., 2007a). The majority of SNPs in *CYP2B6* do not encode functional changes and are relatively rare. However, several important SNPs have been found at higher frequencies, up to 50%, in certain populations. For example, *CYP2B6*6* (516 G>T, 785 G>A) is the most widely studied allele with a proven impact on the pharmacokinetics of *CYP2B6* substrate drugs. Other alleles affecting enzyme function include *CYP2B6*2* (64 C>T), *CYP2B6*3* (777 C>A), *CYP2B6*4* (785 A>G), *CYP2B6*7* (516 G>T, 785 A>G, 1459 C>T) *CYP2B6*16* (785 A>G, 983 T>C), *CYP2B6*18* (983 T>C), *CYP2B6*27* (593 T>C) and *CYP2B6*28* (1132 C>T), as well as four phenotypic null alleles *CYP2B6*8* (415 A>G), *CYP2B6*11* (136 A>G), *CYP2B6*12* (296 G>A) and *CYP2B6*15* (1172 T>A). SNPs in non-coding regions may be in linkage with SNPs in coding regions. As *CYP2B6* is a highly inducible enzyme, SNPs in the promoter 5' UTR may be important for its expression. For example, the recently discovered *CYP2B6*22* (-1848 C>A, -801 G>T, -750T>C, 82 T>C) has been identified in Caucasians and results in enhanced transcriptional activation of up to 9-fold compared with the reference *CYP2B6* promoter. This could cause ultra-rapid metabolism of *CYP2B6* substrates.

A recent study identified the new allele *CYP2B6*29* in a Caucasian individual carrying a heterozygous 68kb deletion of exons 1-4 (Rotger et al., 2007). This adds to the ever-increasing complexity of *CYP2B6*, calling for modifications of genotyping protocols for its alleles.

1.4.4.1.2 CYP2C

The human *CYP2C* subfamily contains four highly homologous genes: *CYP2C8*, *CYP2C9*, *CYP2C18* and *CYP2C19*, which are located in a cluster on chromosome 10 (Gray et al., 1995).

Both the *CYP2C9* and *CYP2C19* genes contain nine exons, encoding proteins of 490 amino acids in length. Although these genes are highly homologous (92%), the enzymes differ in their substrate specificities (Tsao et al., 2001).

CYP2C9 is the main *CYP2C* enzyme, constituting 20% of total human liver microsomal CYP content (Shimada et al., 1994). *CYP2C9* variants *CYP2C9*2* and *CYP2C9*3* are the most common and affect the metabolism of the anti-coagulant drug warfarin. Other drugs affected by *CYP2C9* polymorphism are the anti-diabetic agents glipizide and tolbutamide, the anti-epileptic agent phenytoin, the anti-hypertensive drug losartan and non-steroidal anti-inflammatory drugs (NSAIDs) such as ibuprofen and diclofenac (Miners and Birkett, 1998). Other alleles affecting metabolism of *CYP2C9* substrates are *CYP2C9*5*, *CYP2C9*8*, *CYP2C9*10* and *CYP2C9*11* (Allabi et al., 2004).

CYP2C19 metabolises S-mephenytoin, which is the probe drug for this enzyme. Other drugs metabolised by *CYP2C19* include the anti-ulcer drug omeprazole, diazepam, anti-malarial drug, proguanil and anti-psychotic drug, escitalopram. The most common allelic variants CYP2C19*2 and CYP2C19*3 cause reduced enzyme activity and contribute to the slow metabolism of substrate drugs (Goldstein, 2001), whilst CYP2C19*17 was recently discovered and results in increased enzyme activity (Sim et al., 2006).

1.4.4.1.3 CYP2D

The *CYP2D* gene locus on chromosome 22 contains three contiguous genes: *CYP2D6*, *CYP2D7*, and *CYP2D8* (Kimura et al., 1989b). *CYP2D6* is the only one encoding a functional protein, the others are pseudogenes.

CYP2D6 was the first CYP for which a classical pharmacogenetic polymorphism was found Mahgoub et al., 1977). It remains the most extensively studied and the most polymorphic CYP, with up to 70 alleles described so far, causing a spectrum of phenotypic responses. Besides, CYP2D6 appears to have a large genetic influence on its phenotypes, since environmental factors are of minor importance in CYP2D6 drug oxidation (Bock et al., 1994), which represents a stable and reproducible personal parameter.

Despite the fact that *CYP2D6* is one of the least abundant CYPs in the liver, it metabolises a wide range of drugs, such as anti-arrhythmic agents, tricyclic anti-depressants, neuroleptics and anti-cancer agents (Ingelman-Sundberg, 2005).

Major *CYP2D6* alleles are *CYP2D6*4*, *CYP2D6*10*, *CYP2D6*17* and *CYP2D6*29*. Unequal crossover between *CYP2D6* and the pseudogene *CYP2D7*, involving a certain repetitive sequence, leads to recombination events that result in structural variants in which the functional gene can be either deleted (*CYP2D6*5*) or duplicated (*CYP2D6*1XN* or *CYP2D6*NXn*) (Gaedigk et al., 1991; Ledesma and Agundez, 2005).

1.4.4.2. Flavin mono-oxygenases (FMO)

Members of the microsomal flavin-containing mono-oxygenase (FMO) family catalyse oxidative reactions of large numbers of N-, P- and S-containing xenobiotics.

FMOs and CYPs exhibit similar tissue and cellular locations, molecular weight and substrate specificity (Krueger and Williams, 2005), but often yield distinct metabolites with potentially significant pharmacological consequences. A favourable characteristic of FMOs over CYPs is the lack of reactive oxygen intermediates in its metabolic conversion of xenobiotics (Alfieri et al., 2008; Krueger and Williams, 2005).

The best studied of these enzymes is *FMO3* (in adult human liver), polymorphisms of which contribute to the disease trimethylaminuria, or fish odour syndrome. Although the physiological role of FMO3 is still unclear, knowledge about its metabolism of drugs, xenobiotics and other chemicals is accumulating.

Common non-synonymous SNPs, which reduce enzyme activity, include g.15167 G>A (E158K), g.18281 G>A (V257M) and g.21443 A>G (E308G), with variation found between ethnic groups (Cashman et al., 2000); Mao et al., 2009). Some of these polymorphisms have been associated with a more favourable response to the NSAID sulindac in colorectal cancer patients (Hissamudin et al., 2005).

1.4.4.3. Glutathione S-transferases (GST)

GSTs are dimeric enzymes that catalyse the conjugation of glutathione to electrophilic xenobiotics in order to inactivate them and facilitate their excretion from the body. Such detoxification of potentially

genotoxic compounds gives GSTs an important protective role in chemical carcinogenesis. Various carcinogens, such as the reactive benzo[*a*]pyrene diol-epoxide aflatoxin- 2,3-epoxide and reactive sulfates are GST substrates. However, several chemotherapeutic compounds are also GST substrates, and their conjugation to glutathione might lead to decreased effectiveness as well as tumour cell resistance. In addition, GSTs metabolise certain pesticides and environmental pollutants and have an intracellular transport function.

GST isozymes are divided into seven classes, five of which are cytosolic (alpha, kappa, mu, pi and theta) and two are membrane bound. *GSTMu*, *GSTP* and *GSTT* sub-families appear to be the most polymorphic and give rise to phenotypic variation in their metabolic activity.

GSTMu genes are located in a 20kb cluster on chromosome 1, ordered 5'-*GSTM4-GSTM2-GSTM1-GSTM5-GSTM3*-3' (Pearson et al., 1993). The well characterised *GSTM1* features duplication and deletion (null) variants and SNP alleles *GSTM1*A* and *GSTM1*B*. In addition, a duplicated GSTM1 gene has been associated with ultra-rapid metabolism (McLellan et al., 1997), although this occurs at a low frequency. In the *GSTP* and *GSTT* families, there are *GSTP1* SNP variants such as *GSTP1*A*, *GSTP1*B*, *GSTP1*C* and *GSTP1*D* and a *GSTT1* null allele.

Glutathione S-transferases metabolise environmental toxins, and polymorphisms may play a role in susceptibility to some cancers (Ates et al., 2005b; Mo et al., 2009; Ye and Song, 2005). In addition, these enzymes have also been implicated in increased sensitivity to

chemotherapy due to their role in conjugating drug intermediates (Barahmani et al., 2009; Stoehlmacher et al., 2002). Some GST polymorphisms may influence cancer survival in an indirect way, by detoxifying reactive oxygen species that act as intermediates in the cytotoxicity of chemotherapeutic agents, hence modulating the drug response, even if the compound is not a GST substrate (Ekhart et al., 2009). For example, in individuals with *GSTM1/GSTT1* null genotypes and reduced enzyme activity there will be a higher effective dose of the drug and/or more effective reactive oxidant damage to the tumour tissue. On the other hand, the same patients would be expected to have higher levels of toxicity.

1.4.4.4. N-acetyltransferases (NAT)

The NAT family consists of two isozymes, NAT1 and NAT2. NAT enzymes catalyse the acetylation of a wide variety of arylamines and heterocyclic aromatic amines (Boukouvala and Fakis, 2005). NAT polymorphisms were some of the first genetic variants found in drug metabolism more than 50 years ago (Mitchel and Bell, 1957). Cloning of the N-acetyltransferase 2 (NAT2) cDNA, encoding the enzyme responsible for isonaizid metabolism, was achieved in 1991 by Blum et al., 1991.

The *NAT2* gene is the best characterised, consisting of an intron-less gene of 870 nucleotides encoding a protein of 290 amino acids, with an additional non-coding exon in the 5'-flanking region. Major polymorphisms include the *NAT2*5*, *NAT2*6*, *NAT2*7* and *NAT2*14* alleles. It has been postulated that individuals carrying these alleles metabolise drugs slower and hence are referred to as slow acetylators. An online phenotype prediction apparatus is available to estimate the acetylator status of individuals carrying the common *NAT2* alleles

(Kuznetsov et al., 2009). *NAT2* metabolises the anti-TB drug isoniazid and sulphamethoxazole–trimethoprim (cotrimoxazole) used in HIV patients. Slow acetylators taking *NAT2* substrates such as hypertensive drug hydrazine and anti-arrhythmic agent procainamide are at increased risks for the occurrence of autoimmune disorders (Weinshilboum and Wang, 2006). The hypothesis that acetylator status may predispose to a determined cancer risk is based on a differential effect of N-acetylation as a potential detoxification step and O-acetylation as a potential carcinogen-activation step. For example studies have shown association of *NAT2* polymorphisms with increased risk predisposition to bladder, colon and lung cancer in the background of dietary factors (Lin et al., 2009; Zupa et al., 2009) whilst this is not true for breast cancer. Other studies show that chemicals in smoke may affect acetylator status (Alberg et al., 2004). Therefore NAT2 polymorphism alone does not constitute a relevant risk factor for cancers. However this polymorphism may reinforce the effect of other genetic and/or environmental factors.

1.4.5. Personalised medicine and pharmacogenomics

Pharmacogenomics is based on association studies of genotype and phenotype, aimed at translation of genetic data into clinical practice (Srinivasan et al., 2009). Molecular diagnostics is used to test for variations in genes and their expression, to enable treatment with targeted drugs, thus shaping a narrower, and more potent, definition of personalised medicine (Ferrara, 2007). Pharmacogenomic tests have the potential to (i) predict intended response, the goal outcome of the medication; (ii) predict unintended response to the medication, such as adverse events; (iii) titrate medication dose; and (iv) inform the development of novel therapeutics (Wu and Fuhlbrigge, 2008).

1.4.5.1. Pharmacogenomics in drug development

During early stage clinical trials, pharmacogenomics can be used to improve the safety and efficacy of new drugs in development (Roses, 2004a). Risk of ADRs can be reduced by checking whether a drug will cause problems in genetic subgroups. In reverse, 'failed' drugs can be 'rescued' for application in suitable new target groups. In later stage trials, pharmacogenomics can be applied to screen for 'good responders', increasing confidence by associating response with genotype, thereby reducing trial size and cost (Roses, 2004b). The US Food and Drug Administration (FDA) has recently laid out recommendations for the submission of pharmacogenomic data for new drug approvals (Ruano et al., 2004).

Pharmacogenetics improves the safety of licensed drugs through pre-prescription testing for risk of ADRs, post-marketing surveillance and the use of efficacy data in drug marketing. Examples of the use of pharmacogenetics in the drug discovery pipeline can be found on http://clinicaltrial.gov/ct2/results?term=pharmacogenetics.

1.4.5.2. Pharmacodiagnostic markers

Most established pharmacodiagnostic markers are based on variation in drug metabolising enzyme genes such as CYPs and UGTs, a few drug receptors and transporters (Table 3). Such variation has been observed in the response to a wide range of drugs, illustrated by the following examples.

Warfarin is an anti-thrombotic agent, which is used to prevent strokes. Some individuals suffer from excessive bleeding episodes when administered with this drug. This is largely attributed to polymorphisms in CYP2C9 and the target enzyme, vitamin K epoxide

reductase (VKORC1). On 16 August 2007, the US FDA announced that warfarin's label will carry new information stating that a lower initial warfarin dose "should be considered for patients with certain genetic variations" (Vladutiu, 2008). This marked one of the first official endorsements of genetics' role in drug dosing.

Although genotyping for CYP2C9 and VKORC1 has been widely shown to be of importance in predicting a starting dose of warfarin, a recent study showed that the dosing algorithms are based on common variants only (CYP2C9*2, CYP2C9*3 and VKORC1 1639/3673) and are not predictive for non-Caucasian populations (Langley et al., 2009). Clinical studies indicate that African Americans require higher doses of warfarin, yet the genetic basis for this is not well established (Langley et al., 2009). Therefore, incorporating in dosing algorithms rare variants, which may affect warfarin action in Africans, would be necessary, although small sample sizes might make it difficult to achieve sufficient statistical power.

Anti-epileptic drug phenytoin has a narrow therapeutic index, and toxicity has been associated with polymorphisms in *CYP2C9* and *CYP2C19* genes (Hung et al., 2004). Anti-depressants such as nortriptyline and amitryptyline may require pharmacogenetic testing (Seeringer and Kirchheiner, 2008) to predict individuals who are prone to ADRs.

Cardiovascular drug Bidil is now recommended for Africans, due to reduced side effects compared to Caucasians (Taylor et al., 2004). Although spurring debate around racial issues, this has been endorsed

by FDA recommendations for use of the drug in African Americans (Seeringer and Kirchheiner, 2008).

Breast cancer drug tamoxifen is rendered less effective in Caucasians carrying inactivating *CYP2D6* polymorphisms, which may be an important predictor for the benefits of the drug (Rae et al., 2009).

Camptosar (Irinotecan) is used to treat colorectal cancer, and some patients may suffer from toxicity due to reduced metabolic activity of the enzyme UGT1A1 (Perera et al., 2008).

Some anti-cancer drugs such as 6-mercaptopurine are metabolised by thiopurine S-methyl transferase (TPMT), one of the most important polymorphic enzymes, and poor metabolisers are likely to suffer from increased risk of life-threatening myelosuppression. TPMT is one of the first genes to be tested for, bringing pharmacogenetics 'from the bench to the bedside' (Weinshilboum and Wang, 2006).

Apart from their protective role against environmental toxins, GSTs play a role in cancer treatment. It has been reported that individuals with *GSTM1/GSTT1* null alleles have a better breast cancer prognosis for overall survival when treated with cyclophosphamide (Ambrosone et al., 2001). In contrast, *GSTT1* null genotype was found to be an independent prognostic factor for shorter lung cancer survival.

Anti-retroviral treatment (ART) against HIV/AIDS is known to cause ADRs according to DME-independent genetic variation (Mehta et al., 2008). The HLA-B*5701 allele was shown to predict hypersensitivity reactions after treatment with abacavir (Saag et al., 2008;

Vandekerckhove et al., 2008). However, due to a lower prevalence of the HLA-B*5701 allele in African populations, abacavir hypersensitivity was reported less frequently in African patients than in Caucasians. In addition, studies on the impact of *CYP2B6* polymorphisms on efavirenz treatment show promising results towards targeted application of pharmacogenetics in sub-Saharan Africans (Nyakutira et al., 2007).

In summarry the list of pharmacodiagnostic markers is growing fast, making them an increasingly important pillar of personalised medicine. Pharmacogenetic testing is becoming more widely accepted and partly mandated for therapy decisions and drug development. However, as this trend is highly dependent on diagnostic technology resources, it is so far limited to the more affluent parts of the world. Contributing to its wider application in Africa is the main goal of this study.

Table 3: Examples of pharmacogenetic applications in drug therapy

Disease/ condition	Drugs	Polymorphic Genes	Pharmacogenetic Effect	Reference
Anti-coagulation therapy	Warfarin (Coumadin)***	*CYP2C9, VKORC1*	Haemorrhage in poor metabolisers	(FDA, 2007; Takahashi et al., 2006)
Asthma	Tranilast	*UGT1A1*	Hyperbilirubinemia	(Danoff et al., 2004)
Asthma	Albuterol	*b2-Adrenergic receptor*	Reduced Drug Responsiveness	(Martinez et al., 1997; Tsai et al., 2006)
Breast Cancer	Tamoxifen; Herceptin***	*CYP2D6; ERBB2*	Treatment response	(Bartlett, 2005; FDA, 2008; Goetz et al., 2007)
Cancer	Anthracycline, cyclophosphamide, platinum based	*GSTM1, GSTT1, GSTP1*	Treatment response	(Ambrosone et al., 2001; Stoehlmacher et al., 2002)
Cardiovascular disease	Bidil	-	Improved treatment in Africans	Taylor et al., 2004
Colorectal cancer	Irinotecan (CPT-11; Camptosar)***;	*UGT1A*	Diarrhea, neutropenia	(FDA, 2005; Minami et al., 2007)
Depression	Tricyclic anti-depressants	*CYP2D6*	Adverse reactions	(Seeringer and Kirchheiner, 2008; Steimer et al., 2005)
Epilepsy	phenytoin	*CYP2C9*	Neurotoxicity	(Tate et al., 2005; van der et al., 2001)
Gastrointestinal disease	Omeprazole	*CYP2C19*	Sub-therapeutic drug exposure	(Baldwin et al., 2008)
HIV/AIDS	efavirenz	*CYP2B6, CYP3A4*	Neuropsychosis, hepatotoxicity	(Haas et al., 2004; Rotger et al., 2005)
HIV/AIDS	Abacavir	*HLA-B*	Hypersensitivity reaction	(Mallal et al., 2002)
Inflammatory bowel syndrome	6-Mercaptopurine	*TPMT*	Leukopenia, myelosuppression	(Gisbert et al., 2006)
Malaria	Proguanil	*CYP2C9*	Gastro-intestinal complications	(Kancko et al., 1999)
Malaria	Primaquine	*G6PD*	Haemolytic anaemia	Cordes, 1926; Beutler et al., 2007
Stroke prevention	Plavix (clopidogrel)***	*CYP2C19*	Cardiovascular complications	(FDA, 2009; Simon et al., 2009)
Tuberculosis	Isoniazid	*NAT2, CYP2E1, GSTM1*	hepatotoxicity	(Roy et al., 2008)

***FDA approved pharmacogenetic modifications

1.5. Building African capacity for the future

Capacity building for pharmacogenetics in Africa remains a challenge, struggling with scarce resources and under-representation of African populations in international data repositories (Hardy et al., 2008). As a foundation for future progress, to bridge the gap between research and clinical application and tailor genomic medicine for Africa's health needs, efforts ought to be focused on bioresources and knowledge bases, cataloguing African-relevant information (Singer et al., 2007).

1.5.1. Bioresources for pharmacogenetics research

African biobanks are still few and far between. Africa's biggest biobank in Gambia contains over 50,000 blood samples of individuals suffering from infectious diseases (Sirugo et al., 2004). Other collections have been started in consortia-based efforts (Matimba et al., 2008), focusing on pharmacogenetics in African populations. Given the heterogeneity of the African genome, new and expanded biobanking programmes need to be initiated on the continent.

1.5.2. Pharmacogenomics knowledge in Africa

As the old "one treatment fits all" paradigm becomes less applicable everywhere, significant advances have been made in companion diagnostics and according treatment of patient sub-groups, particularly with newer rationally-targeted therapies (Germano and O'Driscoll, 2009). However, pharmaceuticals developed in Western countries do not yet undergo systematic clinical trials in Africans. Likewise, their effects in African populations are not specifically documented in post marketing surveillance, endangering African patients (Matimba et al., 2008). As a start on the way out of this dilemma, the current status of African pharmacogenetics needs to be assessed to chart future direction.

In order to ensure efficacy of treatment across ethnic groups, the discovery of new pharmacogenetic variants by re-sequencing analysis of African populations, combined with expression profiling and the set-up of knowledge bases on drug response in Africa, ought to be prioritised. The establishment of an information portal enabling the estimation of pharmacogenetic diversity at the population and individual level should create a valuable resource for clinical practice and the pharmaceutical industry.

Therefore, here diverse African ethnicities were analysed using re-sequencing and genotyping strategies in major drug metabolising enzyme genes. The specific aim was to ascertain for polymorphisms in Africans, determining their baseline prevalence, assessing population relatedness and establishing pharmacogenetics resources, such as a catalogue of African-specific pharmacodiagnostic markers. The following aims and objectives of this study were focused on improving the prediction of individual drug response in Africa, as well as the development of novel diagnostics and a better representation of Africans in clinical studies.

2. AIMS AND OBJECTIVES

2.1. AIM

To characterise genetic variation of major drug metabolising enzymes in African populations to understand its impact on biological function, evolutionary relation and medical use as pharmacodiagnostic markers for personalised medicine in Africa.

2.2. OBJECTIVES

1. To collect DNA samples from ethnic groups in Africa including Igbo, Hausa, Yoruba (Nigeria); Kikuyu, Luo, Maasai (Kenya), Shona, San (Zimbabwe), Venda (South Africa) and Mixed Bantu of Tanzania.

2. (i) To search for novel genetic variants in genes of major drug metabolising enzymes, namely *CYP2C9*, *CYP2C19*, *CYP2D6* and *NAT2* by re-sequencing.
   (ii) To predict functional characteristics of these variants using bioinformatic tools.

3. To investigate the baseline frequencies of commonly known alleles in eight genes, namely *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *NAT2*, *GSTM1*, *GSTT1* and *FMO3* by genotyping.

4. To assess population relatedness of Africans by applying statistical and phylogenetic methods to genotype data.

5. To establish a catalogue of polymorphisms in drug metabolising enzyme genes towards a pharmacogenetics database of African populations.

3. MATERIALS AND METHODS

3.1. Populations

3.1.1. Ethics

The study was based on the Declaration of Helsinki of the World Medical Association (www.wma.net/e/policy/b3.htm). Accordingly, research proposals regarding the 'collection of blood samples for the analysis of genetic variability of ADME genes' were approved by the Ethics Review Boards of University of Nairobi (Kenya), Obafemi Awolowo University (Nigeria) and University of Zimbabwe (Zimbabwe). Samples from South Africa and Tanzania were analysed on the basis of a previously approved agreement for the study of pharmacogenetics (Dandara et al., 2002, PhD thesis). Overall, the study was part of an effort by the Consortium of Study for Pharmacogenetics of African Populations (Matimba et al., 2008).

3.1.2. Population selection and volunteers

Populations were selected based on ethno-linguistic classifications (www.ethnologue.org). Ethnicity was assigned based on the submission that parents and grandparents of the volunteers were of the same self-identified ethnic group. The study subjects consisted mainly of unrelated university students, or of adults from isolated villages. The volunteers were educated on the nature of the study, in the appropriate language, confirmed agreement and understanding by signing informed consent forms. In some cases, such as the Hausa in Nigeria, additional permission was obtained from the Chiefs of the communities. All samples were anonymised and renamed to fit the database coding system, which was based on the country, ethnicity

and a number. All samples were re-coded such that there was no direct link to the original code given by the collector.

3.1.3. Blood samples

Blood samples from the following ethnic groups were analysed (number of samples): Kikuyu (100), Luo (100), Maasai (152), Hausa (100), Igbo (100), Yoruba (100), San (64), Shona (100), Mixed Tanzanians (100), Venda (81). 5 to 10ml venous blood were collected in EDTA-containing Venoject vacutainer tubes (Terumo Europe, Leuven, Belgium) and stored at -20°C.

3.1.4. DNA extraction and quality control

DNA was extracted using the Eppendorf Perfect gDNA Blood Mini Kit (Eppendorf, Hamburg, Germany) or the QIAamp DNA Blood Mini Kit (Qiagen, Venlo, Netherlands). 20µl of DNA were diluted 1:20 in distilled water, and absorbance was measured at 260/280nm on a UV/2101PC Spectrophotometer (Shimadzu, Kyoto, Japan). For quality control, extracted genomic DNA was checked by electrophoresis in 1% agarose Roche Applied Science (Mannheim, Germany) against molecular weight marker MWVIII (Roche, Basel, Switzerland). Samples that produced bands of about 20kb and above were stored, while degraded samples were re-extracted.

3.2. Genes

3.2.1. Selected genes

Eight drug metabolising enzyme genes were selected for this study: *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *FMO3*, *GSTM1*, *GSTT1*, *NAT2*. Their accession IDs were obtained from http://genome.ucsc.edu/ and are shown in Table 4, unless otherwise stated. SNPs or CNVs were

analysed using PCR, sequencing and genotyping methods such as RFLP, Taqman and Sequenom (Table 4).

Table 4: Genes analysed in this study

Gene	Accession ID‡	Variants¶	Study methods
CYP2B6	NM_000767	SNPs	PCR-RFLP, Sequenom
CYP2C9	NM_000771 (NC_000010.9)	SNPs	Re-sequencing, PCR-RFLP, Sequenom
CYP2C19	NM_000769 (NC_000010.9)	SNPs	Re-sequencing, PCR-RFLP, Sequenom
CYP2D6	M33388	CNV, SNPs, indels	Re-sequencing, Long Range PCR, Nested PCR-RFLP
FMO3	NM_006894	SNPs	Taqman
GSTM1	NT_019273.18	CNV	PCR
GSTT1	NT_011520.11	CNV	PCR
NAT2	NM_000015 (NC_000008.9)	SNPs	Re-sequencing, PCR-RFLP, Sequenom

‡Accession ID according to http://genome.ucsc.edu/ (in brackets: Alternative accession reference sequences obtained from www.ncbi.nlm.nih.gov/gene)
¶Major variant types important in pharmacogenetics

3.2.2. Allele nomenclature

Allele nomenclature of CYPs and *NAT2* is according to the Human Cytochrome P450 (CYP) Allele Nomenclature Committee (Ingelman-Sundberg et al., 2001) (www.cypalleles.ki.se) and the Arylamine N-acetyltransferase Gene Nomenclature Committee (Hein et al., 2008) (http://louisville.edu/medschool/pharmacology/nat), respectively.

An allele, as borne by a specific SNP, is assigned a number preceded by a star - for example, *CYP2C9*9*, to define the 10535 A>G mutation, causing the amino acid change H251R. For FMO3, variants are defined by the result of the amino acid change and its position - for example, the SNP g.15167 G>A (E158K) is referred to as the K158 variant.

3.3. Re-sequencing

The following numbers of DNA samples from various ethnicities were analysed: Hausa (20), Igbo (20), Luo (30), Maasai (13), San (40), Shona (23), Venda (9), Yoruba (20) and Tanzanian Mixed Bantu (12).

Primers were designed using SNPBox and Primer3 software (Rozen and Skaletsky, 2000; Weckx et al., 2004). Identical primers were used for PCR and sequencing, except where otherwise stated (Table 5).

Exon amplification mixtures (20µl) contained 1xTiTaq buffer, 0.25mM of deoxyribonucleotide triphosphates, dNTP (dATP, dCTP, dGTP, dTTP), 0.5µM of each primer, 0.25 units of TiTaq and 5ng/µl of DNA template. 7-Deaza-2'-deoxyguanosine-GTP (7-deaza-dGTP) was used in place of dGTP in CYP2D6 reactions. All reagents were obtained from Invitrogen SA (Merelbeke, Belgium). For PCR, an initial denaturation at 94°C for two minutes was followed by 35 cycles at 94°C for 30 seconds, 59°C for 30 seconds and 72°C for 60 seconds. PCR product were purified using ExoSAP-IT® reagent (USB Biotechnologies, Cleveland, USA). Sequencing mixes (15 µl) contained 2 µl of cleaned-up PCR product (5x diluted), 0.5 µM of sequencing primer and 0.6 µl of Big Dye® Terminator Sequencing mix (Applied Biosystems, Foster City, CA). Sequencing reactions were started at 96°C for one minute followed by 25 cycles at 96°C for ten seconds, 50°C for five seconds and 60°C for four minutes, and resolved on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, CA). Identification of SNPs was carried out using the novoSNP v2.1.9 software package (Weckx et al., 2005). Reference sequences were NC_000010.9 for *CYP2C9*, NC_000010.9 for *CYP2C19*, M33388 for *CYP2D6* and NC_000008.9 for *NAT2*.

Table 5: Exon amplification and sequencing primers

Gene	Exon	Exon amplification primers	Sequencing primers
CYP2C9	5'UTR	**cyp2C9-**5'FLF ATCCTCAACTCAGTATGTCAGC **cyp2C9-**5'FLR ATCACCTAGGTCCACTATATGC	**cyp2C9-**5'FLSF1 ATCCTCAACTCAGTATGTCAGC **cyp2C9-**5'FLSR1 ACCTTTACCATTAAACCCCC **cyp2C9-**5'FLSF2 CAATTCCTGCCTTCAGGA **cyp2C9-**5'FLSR2 AAGGACTTTGACCCACTGAT
	1	**cyp2C9-1F** GGAATGTACAGAGTGGACAATGG **cyp2C9-1R** GATCCCACAATACCTTACCATTTAC	***
	2&3	**cyp2C9-2&3F** GACCTGCTGAATATGTTGATGTG **cyp2C9-2&3R** CCCGCTTCACATGAGCTAAC	**cyp2C9-**2SF TCTTGAACTCCTGACCTTGT **cyp2C9-**2SR GGAGCTCTGTAAGTCTCTGT **cyp2C9-**3SF AGGAGTTTTCTGGAAGAGG **cyp2C9-**3SR GGAAAAACACTGCTCTTTAACTC
	4*	**cyp2C9-4F** CAGCTAGGTTGTAATGGTCAACTC **cyp2C9-4R** GCTAATGGGCTTAGAAATCAGG	***
	5*	**cyp2C9-5F** TCATCTGGTTAGAATTGATCCTCTG **cyp2C9-5R** GCTATTAACTACCGCCTCAACTTC	***
	6*	**cyp2C9-6F** GAGGAAATGGACCTAGAGACCTTC **cyp2C9-6R** CCCATTGTAATCACCATTAGTTTG	***
	7*	**cyp2C9-7F** GTGCATCTGTAACCATCCTCTCT **cyp2C9-7R** CAGACACTAGGACCTGTTACAAACC	***
	8*	**cyp2C9-8F** AGAAGGTTGCATCCAAGTATCC **cyp2C9-8R** GAGTTCTTGGGTACCTCACTGGT	***
	9	**cyp2c9-9F** CTCATCCATCCATTCATTCATG **cyp2c9-9R** CTCTAACACTCACCCAAAATAGC	**cyp2c9-9SF** CTCATCCATCCATTCATTCATG **cyp2c9-9SR** CGAATGTTCACTAGATCTTCAG **cyp2c9-9S2F** CTGCAGCTCTCTTTCCTC **cyp2c9-9S2FR** CTCTAACACTCACCCAAAATAGC
CYP2C19	1	**cyp2C19-1F** CAATTATGACGGTGCATTGG **cyp2C19-1R** CACTTCCCTTACTGTTTACCCTCA	***
	2&3	**cyp2C19-2&3F** GTTCTTGAAGCTGGGTATTTGTC **cyp2C19-2&3R** AGCAAAGTTCAGGAGAACATAGG	**cyp2C19-2SF** AATTCAGAAATATTTGAGCCTGTGTG **cyp2C19-2SR** GGTTTTTCTCAACTCCTCCACAA **cyp2C19-3SF** GCCTGGGATCTCCCTCCTAGTTT **cyp2C19-2&3R** AAGCAAAGTTCAGGAGAACATAGG
	4	**cyp2C19-4F** CAGCTAGGCTGTAATTGTTAATTCG **cyp2C19-4R** GAGTAATGGAAGACTCCAAAGTGC	***
	5	**cyp2C19-5F** TTCAATTTCAGAGGCTGCTTG **cyp2C19-5R** CTATGATGCTTACTGGATATTCATGC	***

Table 5 continued: Exon amplification and sequencing primers

Gene	Exon	Exon amplification primers	Sequencing primers
CYP2C19	6	**cyp2C19-6F** CAGCATATAAACAGAGCCAAAGAC **cyp2C19-6R** ACACCATTAAATTGGGACAGATTAC	***
	7	**cyp2C19-7F** CCTAGCTTAAGGCACAGTTACACA **cyp2C19-7R** GAAAGACTCAAGGTGTCAAGATGTC	***
	8	**cyp2C19-8F** GCCTTAAGCTCATGCCTCTTATTAC **cyp2C19-8R** GGCAGAATTCAACCAACCTATACTT	***
	9	**cyp2C19-9F** TCATTGTTTAGTTGCCTATCCATC **cyp2C19-9R** CCATCTTCACCTTTGTCCTTTC	***
CYP2D6	1&2	**cyp2D6ex1_2F** ACCAGGCCCCTCCACCGG **cyp2D6ex1_2R** CTCTCTGCCCAGCTCGG	**CYP2D6ex1_2F** ACCAGGCCCCTCCACCGG **cyp2D6ex1SR** GTTTCACCCACCACCCATGTTT **cyp2D6ex2SF** CTTCCACCTGCTCACTCCTGGTA **cyp2D6ex2SR** CCTCCCTAGTGCAGGTGGTTTCT
	3&4	**cyp2D6ex3_4F** ATTTCCCAGCTGGAATCC **cyp2D6ex3_4R** GAGACTCCTCGGTCTCTC	**cyp2D6ex3_4SF** GAGCATAGGGTTGGAGTGGGTG **cyp2D6ex3_4R** GAGACTCCTCGGTCTCTC
	5&6	**cyp2D6ex5_6F** GCCTGAGACTTGTCCAGG **cyp2D6ex5_6R** CCGGCCCTGACACTCCTTCT	**cyp2D6ex5_6F** GCCTGAGACTTGTCCAGG **cyp2D6ex5_6R** CCGGCCCTGACACTCCTTCT
	7,8,9	**cyp2D6ex7_9F** GGATCCTGTAAGCCTGACCTC **cyp2D6ex7_9R** ACTGAGCCCTGGGAGGTAGGTAG	**cyp2D6ex7_9F** GGATCCTGTAAGCCTGACCTC **cyp2D6ex7SR** GTGGTGGCATTGAGGACTAGGTG **cyp2D6ex8SF** GTCCAGAGTATAGGCAGGGCTGG **cyp2D6ex8SR** AGCACAAAGCTCATAGGGGGATG **cyp2D6ex9SF** CTTCCTCTTCTTCACCTCCCTGC **cyp2D6ex9SR** AATATGGGCCTCCAGGCTGAGT
NAT2	2	**NAT2ex2F** GAAGCATATTTTGAAAGAATTGG **NAT2ex2R** GCATTTTAAGGATGGCCTGT	**NAT2ex2F** GAAGCATATTTTGAAAGAATTGG **NAT2SF1** TGCCAAAGAAGAAACACCAA **NAT2SR2** ACCTCGAACAATTGAAGATTTTGA **NAT2ex2R** GCATTTTAAGGATGGCCTGT

***same set of primers used for sequencing.

3.4. Genotyping

3.4.1. RFLP

RFLP genotyping was according to previously established methods (Abe et al., 1993; Bell et al., 1993; Brockmoller et al., 1992; Gaedigk et al., 1999; Pemble et al., 1994; Rotger et al., 2005), optimised as detailed in Table 6. The following alleles were analysed: *CYP2B6*6*, *CYP2C9*2*, *CYP2C9*3*, *CYP2C19*2*, *CYP2C19*3*, *CYP2D6*2*, *CYP2D6*4*, *CYP2D6*5*, *CYP2D6*17*, *CYP2D6*29*, *NAT2*5*, *NAT2*6*, and *NAT2*7*.

Primers were purchased from Eurogentec S.A. (Seraing, Belgium). Three Taq polymerases were used in this study: Taq DNA polymerase (Roche Applied Science, Mannheim, Germany), JumpStart REDTaq DNA polymerase and JumpStart REDAccuTaq LA DNA polymerase (Sigma-Aldrich, St. Louis, MO). Restriction enzymes were purchased from New England Biolabs Inc. (Ipswich, MA). dNTPs and gel electrophoresis reagents were purchased from Roche Applied Science (Mannheim, Germany) and Invitrogen SA (Merelbeke, Belgium). For PCR, the MG96G thermal cycler (LongGene Scientific Instruments, Hangzhou, China) and the PTC-100TM Programmable Thermal Controller (MJ Research Inc, Waltham, MA) were used. The conditions for amplification and restriction enzyme digestion are shown in Table 6. Fragments of less than 1kb were separated in 2-3% agarose gels. Large fragments (over 1 kb) were run in 1% gels. Molecular weight markers 1kb or 100bp ladders (Invitrogen SA, Merelbeke, Belgium) were used, respectively. Ethidium bromide (SIGMA, Stockholm, Sweden) was used for visualisation of fragments on a Syngene gel documentation system (Synoptics Ltd, Cambridge, UK)

The *CYP2D6* gene was first isolated from the *CYP2D7* and *CYP2D8* pseudogenes by long-range PCR. This was followed by nested PCRs with primers for the specific SNPs to be analysed (*CYP2D6*4*, *CYP2D6*17*, *CYP2D6*29*).

CNV polymorphisms were detcted as follows: *CYP2D6*5* was identified by multiplex long-range PCR using primers for the whole *CYP2D6* gene and primers specific for *CYP2D6*5*, producing bands of 6.6kb and 3.5kb, respectively. *GSTM1*0* and *GSTT1*0* (homozygous deletion alleles only) (homozygous deletions) were detected by a single PCR reaction only.

3.4.2. Taqman

DNA samples of the following 863 subjects were analysed for *FMO3*: 97 Kikuyu, 99 Luo, and 143 Maasai from Kenya, 99 Shona and 63 San from Zimbabwe, 63 Venda from South Africa, and 100 Hausa, 99 Igbo and 100 Yoruba from Nigeria. Genotyping was carried out by TaqMan allelic discrimination with fluorogenic 5′ nuclease assays in an ABI Prism 7000 Sequence Detection System (Applied Biosystems, Foster City, CA). SNPs were analysed using the following validated TaqMan Genotyping Assays purchased from Applied Biosystems: for D132H, rs12072582, assay ID: C__30633935_10; for E158K, rs2266782, assay ID: C__2461179_30; for V257M, rs1736557, assay ID: C__8698544_30; for E308G, rs2266780, assay ID: C__2220257_30; for L360P, rs28363581, assay ID: C__30633936_20), according to the guidelines of the manufacturer.

Table 6: Alleles, primers, PCR conditions, restriction enzymes (RE) and expected fragment pattern of RFLP analysis

Allele	Primers	PCR conditions	RE	Agarose gel fragment pattern
CYP2B6*6 (516 G>T)	*F* GGTCTGCCCATCTATAAAC *R* CTGATTCTTCACATGTCTGCG	35 cycles of 94°C 20s, 60 °C 20s, 72°C 60s	*Bsr*I	3% w/v gel 216+237bp (wt) 526bp (mt)
CYP2C9*2 (430 C>T)	*F* TACAAATACAATGAAAATATCATG *R* CTAACAACCAGACTCATAATG	35 cycles of 94°C 20s, 56 °C 20s, 72°C 40s	*Ava*II	2% w/v gel 527 + 164bp (wt) 691bp (mt)
CYP2C9*3 (1075 G>A)	*F* AATAATAATATGCACGAGGTCCAGAGGTAC *R* GATACTATGAATTTGGGACTTTC	35 cycles of 94°C 20s, 57 °C 20s, 72°C 40s	*Kpn*I	3% w/v gel 130,35 (wt) 165bp (mt)
CYP2C19*2 (681 G>A)	*F* AATTACAACCAGAGCTTGGC *R* TATCACTTTCCATAAAAGCAAG	35 cycles of 94°C 10s, 53 °C 10s, 72°C 20s	*Sma*I	3% w/v gel 120 +49bp (wt) 169bp (mt)
CYP2C19*3 (636 G>A)	*F* TATTATTATCTGTTAACTAATATGA *R* ACTTCAGGGCTTGGTCAATA-3	35 cycles of 94°C 10s, 53°C 10s, 72°C 20s	*Bam*HI	3% w/v gel 135 + 30bp (wt) 165bp (mt)
CYP2D6 (whole gene)	*F* CCAGAAGGCTTTGCAGGCTTCAG *R* ACTGAGCCCTGGGAGGTAGGTAG	35 cycles of 94°C 20s, 53°C 10 20s, 72°C 8 min	-	1% w/v gel 6.6kb
CYP2D6*4 (1846 G>A)	*F* AGAGGCGCTTCTCCGTGTCCA *R* CAGAGACTCCTCGGTCTCTCG	35 cycles of 94°C 10s, 58 °C 10s, 72°C 20s	*Bst*NI	3% w/v gel 194+161+37bp (wt) 355 + 37bp (mt)
CYP2D6*5 *(deletion of CYP2D6)*	F CACCAGGCACCTGTACTCCTCAG *R* CAGGCATGAGCTAAGGCACCCAGAC	35 cycles of 94°C 20s, 53°C 10 20s, 72°C 8 min	-	1% w/v gel 3.5kb
CYP2D6*17 (1023 C>T)	*F* GTCGTGCTCAATGGGCTGGCGGCCGTGCGCGAGGCG *R* GGCGAGGACACCGCCGACCGCCCGCCTGTGCCC**Agt A**	35 cycles of 94°C 10s, 58°C 10s, 72°C 20s	*Fok*I	3% w/v gel 180 + 74bp (wt) 254bp (mt)
CYP2D6*29 (1659 G>A)	*F* TATGGGCCAGCGTGGAGCGAGCAGAGGCGCT**TCcgC** *R* AGATGCGGGTAAGGGGTCGCCTTCC	35 cycles of 94°C 10s, 58°C 10s, 72°C 20s	*Bst*UI	3% w/v gel 178 + 35bp (wt) 213bp (mt)

Table 6 continued:
Alleles, primers, PCR conditions, restriction enzymes (RE) and expected fragment pattern of RFLP analysis

Allele	Primers	PCR conditions	RE	Agarose gel fragment pattern
GSTM1*0[‡] (gene deletion)	*F* CTGCCCTACTTGATTGATGGG *R* CTGGATTGTAGCAGATCATGC Beta actin internal control primers: *F* TGACGGGGTCACCCACACTGTGCCCATCTA *R* TAGAAGCATTTGCGGTGGACGATGGAGGG	35 cycles of 94°C 10s, 53 °C 10s, 72°C 20s	-	2% w/v gel 273bp for *GSTM1* and 600bp for Beta actin control
GSTT1*0[‡] (gene deletion)	*F* TTCCTTACTGGTCCTCACATCTC *R* TCACCGGATCATGGCCAGCA	35 cycles of 94°C 10s, 56°C 10s, 72°C 20s	-	2% w/v gel 650bp for *GSTT1* gene
NAT2*5 (341 T>C)	*F* TGACGGCAGGAATTACATTGT *R* CCTTGTTTTATTTGGGAACACA	35 cycles of 94°C 20s, 55°C 20s, 72°C 60s	*Asp*718	2% w/v gel 419 +140bp (wt) 559bp (mt)
NAT2*6 (590 G>A)	*F* TGA CGG CAG GAATTACATTGT *R* CCTTGTTTTATTTGGGAACACA	35 cycles of 94°C 20s, 55°C 20s, 72°C 60s	*Taq*I	2% w/v gel 20+143+169+227bp (wt) 396bp (mt)
NAT2*7 (857 G>A)	*F* TGA CGG CAG GAATTACATTGT *R* CCTTGTTTTATTTGGGAACACA	35 cycles of 94°C 20s, 55°C 20s, 72°C 60s	*BamH*I	2% w/v gel 515+44bp (wt) 559bp (mt)

F=forward; R=reverse; bold nucleotides in primers are modifications introducing a restriction site; s=seconds, min=minutes; RE= restriction enzyme; w/v=weight/volume. [†]nested PCR after long-range PCR amplification of whole *CYP2D6* gene; [¶]multiplex with *CYP2D6* whole gene; [‡]analysis for homozygous deletion genotype; Position of nucleotide change is shown in brackets according to reference sequences in Table 4. wt=wild type allele (or original nucleotide); mt=mutant allele (nucleotide change), e.g. 516 G>T indicates nucleotide change from G to T at position 516.

Table 7: Primers for analysis using Sequenom MassARRAY

P	*Gene* SNP ID	Forward Primer	Reverse Primer	Extension primer	EXT: M1	EXT2:M2
1	*NAT2*rs4646247	ACGTTGGATGCAGATCTACACAATAAAGCTC	ACGTTGGATGTCATCCTCTTTCTTGTCTCC	AAGCTCATGTTCCTTCT	G: 5358.5	A: 5438.4
	*NAT2*rs1801280	ACGTTGGATGATACAGCACTGGCATGGTTC	ACGTTGGATGCACATCTGGGAGGAGCTTC	TCTCCTGCAGGTGACCA	C: 5393.5	T: 5473.5
	*CYP2B6*rs3786547	ACGTTGGATGTCCTTGTCTGGAAGTGTGAG	ACGTTGGATGTCCTTACTGAGCCTATGTCC	GAGGGAGAGCTCAGCCAC	T: 5829.8	C: 5845.8
	*NAT2*rs721399	ACGTTGGATGGTTCTTAACACTATTCTCCC	ACGTTGGATGTCTAATGTCTGCTTTTCTC	CTATTCTCCCTCTTCGACT	G: 5887.9	A: 5967.8
	*NAT2*rs7832071	ACGTTGGATGGGATTTCCAACTCCTCATGC	ACGTTGGATGCTCTTAGTAGTCTCAGAACC	ACTCCTCATGCTTAAAAGA	C: 6003	T: 6082.9
	*CYP2C19*rs4986894	ACGTTGGATGGTGGGTCAAAGTCCTTTCAG	ACGTTGGATGTAAGTGTGTGTGCCTCTTTG	CCTAGGTGATTGGCCACTT	C: 6042	T: 6121.9
	*CYP2C19*rs4244285	ACGTTGGATGCACTTTCCATAAAAGCAAGG	ACGTTGGATGGCAATAATTTTCCCACTATC	AGTAATTTGTTATGGGTTCC	G: 6400.2	A: 6480.1
	*CYP2C19*rs12779363	ACGTTGGATGTCCCTTTGTATTTCTGGCTC	ACGTTGGATGGGTAACATGTTTAGACATGTG	CAGCCAAAGATTTTTTTCCCG	G: 6603.3	A: 6683.2
	*CYP2C19*rs4304697	ACGTTGGATGTCCTAACCAGCTGTCTCATC	ACGTTGGATGCAGAAGCTGCCAAGAAACAC	TGAGACATCAAACACCTCTGCC	G: 6895.5	A: 6975.4
	*NAT2*rs4646246	ACGTTGGATGTGTGGCATGGGTATCATCAG	ACGTTGGATGTAGTTGCAGGGCCTCAGGTT	AAAAGATTTGCGTAAGAGATTC	A: 7077.7	G: 7093.7
	*CYP2C19*rs4986893	ACGTTGGATGGACTGTAAGTGGTTTCTCAG	ACGTTGGATGAACATCAGGATTGTAAGCAC	AAAAAACTTGGCCTTACCTGGAT	G: 7278.8	A: 7358.7
	*NAT2*rs1799930	ACGTTGGATGCCTGCCAAAGAAGAAACACC	ACGTTGGATGACGTCTGCAGGTATGTATTC	AATATACTTATTTACGCTTGAACCTC	A: 8150.4	G:8166.4
	*CYP2B6*rs8192712	ACGTTGGATGGTCACCTACTGGCATTAGAG	ACGTTGGATGTTTCCCACCTTGTTCAGGAG	ATCAGTTAGATTTGTTTACCCATAAG	C: 8206.4	T: 8286.3
	*NAT2*rs1799931	ACGTTGGATGCACAAGGGTTTATTTTGTTCC	ACGTTGGATGGAGAAATCTCGTGCCCAAAC	GTTCCTTATTCTAAATAGTAAGGGAT	G: 8246.4	A: 8326.3
2	*CYP2B6*rs3745274	ACGTTGGATGTTCAGTCTGTGTCCTTGACC	ACGTTGGATGATGGAGCAGATGATGTTGGC	CCCACCTTCCTCTTCCA	G: 5264.5	T: 5304.3
	*CYP2B6*rs7259965	ACGTTGGATGACTGATGCGTATGAGACAGG	ACGTTGGATGGTGGTTTAGAATCTACTGGG	GGTCTATGGCTTCCACT	T: 5423.6	C: 5439.6
	*CYP2B6*rs2306606	ACGTTGGATGTATTCTGGGAGCACTGTAGG	ACGTTGGATGGAGAACGCATGGACAGATTC	CTCTGGGCTAGATTCCAA	C: 5721.8	T: 5801.7
	*CYP2B6*rs8100458	ACGTTGGATGGACCTGTGAGGAGAAAGAAC	ACGTTGGATGTGACCCTCTTTTCCAACACC	AGTAAGCAAACCTCAAGA	C: 5756.8	T: 5836.7
	*CYP2B6*rs12721649	ACGTTGGATGTCCACAGTAGGGAAGGAAGG	ACGTTGGATGTCAAAGACCTTTAGGCCAAC	CTGTAGCCTTGCTGTGTATA	G: 6361.2	A: 6441.1
	*CYP2C9*rs1057910	ACGTTGGATGTGTCACAGGTCACTGCATGG	ACGTTGGATGCTACACAGATGCTGTGGTGC	GCTGGTGGGGAGAAGGTCAA	C: 6574.3	A: 6614.2
	*CYP2C9*rs9332168	ACGTTGGATGTGATTCAGGCTCACAGTGTC	ACGTTGGATGACCTGTAGACAACCACAATC	TAATGATATGGCCCCAATTTA	C: 6651.4	T: 6731.3
	*CYP2C9*rs9332114	ACGTTGGATGGCTATAAACTCTGTTAGGG	ACGTTGGATGTAGTCCTCACAGCAGTACTC	CTCTGTTAGGGGGTTAAGAAT	C: 6763.4	T: 6843.3
	*NAT2*rs721398	ACGTTGGATGCTGCAGTTTAAACTATAGCCC	ACGTTGGATGGAGAGTGTGCCTTGTGATTC	GTTTAAACTATAGCCCTGATAC	C: 6940.6	T: 7020.5
	*CYP2B6*rs10426235	ACGTTGGATGCCACTCTCCAAACTGAGATG	ACGTTGGATGCAAGATGGGCATAGAGGAG	ACAAAAACGCTTAATGGAACATCA	G: 7594	A: 7673.9
	*CYP2C19*rs4388808	ACGTTGGATGTGATTGGAAAGAGATAAGC	ACGTTGGATGATAACTGTAATTTTATTCC	AAGCAAAATGAAACAGAGCCAAATTA	G: 8269.5	A: 8349.4
	*CYP2C19*rs7088784	ACGTTGGATGCATGTGTAGTAATTCTCTG	ACGTTGGATGAAAGCAGGTATAAGTCTAGG	GAAAATTCTGACAAGAGAATCAAAGA	G: 8300.5	A: 8380.4
3	*CYP2B6*rs2279342	ACGTTGGATGAAGGCTCTGAGTCCTTCTTC	ACGTTGGATGGGA ATGGTGAGAGCCTCAG	TCCTTCTTCCTCCCATA	T: 5278.5	A: 5334.4
	*CYP2B6*rs2279343	ACGTTGGATGCTTTTTCCATGTGGAGCAGG	ACGTTGGATGGTGGAGAAGCACCGTGAAAC	TAGGTGTCGATGAGGTCC	G: 5817.8	A: 5897.7
	*NAT2*rs1799929	ACGTTGGATGTGCTTGACAGAAGAGAGAGG	ACGTTGGATGGGCAGGAGATGAGAATTAAG	GAGAGAGGAATCTGGTAC	C: 5859.9	T: 5939.8
	*CYP2C19*rs11597626	ACGTTGGATGGGTTTAAATTGGAAAAGAGC	ACGTTGGATGGCCTTCTATCACATGGTTTC	TTGGAAAAGAGCAGGTTTGA	G: 6492.3	C: 6532.3
	*NAT2*rs1208	ACGTTGGATGAACTCTCACTGAGGAAGAGG	ACGTTGGATGTTTGGGCACGAGATTTCTCC	AAGAGGTTGAAGAAGTGCTGA	A: 6854.5	G: 6870.5
	*CYP2C19*rs10509676	ACGTTGGATGCGCTTATCTGCTTTTGCTGG	ACGTTGGATGGCCGTGAAAATTGAACTTCTC	CCTTTACAACAGGCATAAATTA	T: 6957.6	A: 7013.5
	*CYP2B6*rs8192711	ACGTTGGATGACCTCCCATAGTTAAAAGCC	ACGTTGGATGACCTTTTCTGAACAGAAACC	GTCAATAACTGAAATTACTATCAA	G: 7575	A: 7654.9
4	*NAT2*rs1801279	ACGTTGGATGTTGATTGACCTGGAGACACC	ACGTTGGATGCCATGGAGTTGGGCTTAGAG	GGAGACACCACCCACCC	G: 5341.5	A: 5421.4

P=multiplexing; EXT=extension nucleotide; M1=mass of extended fragment 1; M2=mass of extended fragment 2.

3.4.3. Sequenom

DNA samples of Hausa (48), Masaai (51), San (40) and Shona (46) were genotyped using the Sequenom MassARRAY system (Sequenom, San Diego, CA). Expecting guidance by HapMap data in tagSNP selection for major haplotypes, SNPs in *CYP2B6*, *CYP2C9*, *CYP2C19* and *NAT2* gene regions were selected from the Yoruba (YRI) HapMap Data Release 21, July 2006 (www.hapmap.org). Haploview v3.31 (Barrett et al., 2005a) was used to mask repeat regions and select SNPs in major haplotype blocks represented at a frequency of 5% and higher, amongst others (Table 7).

PCR primers were designed as sets of forward and reverse primers, together with an extension primer, and reactions were multiplexed, using the Sequenom MassARRAY Assay Design software (Table 7).

PCR mixtures (10µl) contained 1xTiTaq buffer, 0.5mM dNTPs, primer pools of 0.5µM each of forward and reverse primer sets, 1 unit of TiTaq and 20ng of template DNA. All reagents were obtained from Invitrogen SA (Merelbeke, Belgium). PCR conditions were: 35 cycles of 95°C for 20 seconds, 56 °C for 30 seconds and 72 °C for 1 minute. PCR products were purified using arctic shrimp alkaline phosphatase (ExoSAP-IT, USB, Cleveland, OH). The iPLEX gold method (Sequenom, San Diego, CA) was used for single-base primer extension, employing a universal mix of mass-modified ddNTPs, extension primers and DNA polymerase and producing two products of different mass for every SNP (Table 7).

Product mixtures were dispensed on SpectroCHIP Arrays and analysed by MALDI-TOF on the MassARRAY Nanodispenser (Sequenom, San

Diego, CA). Genotypes were inferred based on molecular weight (MW) of oligonucleotides, and Typer v3.3 software was used for further analysis.

3.5. Collection of reference data

Published articles containing frequency data of known polymorphisms were obtained from the PubMed database provided by NCBI (www.pubmed.gov). The following keywords were used: *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *NAT2*, *GSTM1*, *GSTT1*, polymorphism, genetic variation, populations, African, Caucasian, Asian, pharmacogenetics.

3.6. Data analysis

3.6.1. Nucleotide sequence

All genomic sequences were based on reference sequences in NCBI databases (http://blast.ncbi.nlm.nih.gov/Blast.cgi), unless otherwise stated.

The specificity of primers was confirmed by BLAST (Basic Local Alignment Search Tool) analysis with human genomic sequences in NCBI databases.

All novel SNPs, detected by re-sequencing, were verified against the NCBI Single Nucleotide Polymorphism database (dbSNP, www.ncbi.nlm.nih.gov/SNP) using BLAST analysis. Their novelty was confirmed by comparison with the CYP and *NAT2* Allele/Gene

Nomenclature Committee pages online (see 3.2.2), in addition to literature searches.

3.6.2. Haplotype determination

Haploview v3.31 was used to determine haplotypes from genotype data obtained by re-sequencing. LD plots were generated to assess the extent to which SNPs are likely to be linked and hence likely to occur on the same haplotype with logarithm of odds (LOD) score >3. The Haploview Tagger tool was used to estimate which SNPs are likely to be tagged by a single SNP in a predicted haplotype (threshold r^2 >0.8).

3.6.3. Prediction of functional effects of non-synonymous SNPs

Functional effects of non-synonymous SNPs were predicted using the Polyphen prediction programme (http://genetics.bwh.harvard.edu/pph/) (Ramensky et al., 2002), based on position-specific independent count (PSIC) scores of multiple sequence alignments, and structural information, if available. The programme predicts the functional effects of SNPs, based on occurrence in active/binding sites or in transmembrane regions, interference with disulphide or other bonds, compatibility with homologous sequences at that position, as well as mapping to known three-dimensional protein structures or validated homology models. Protein sequence accession numbers were obtained from Swiss-Prot (The Uniprot Consortium, 2008) (www.uniprot.org) as CYP2C9: P11712; CYP2C19: P33261; CYP2D6: P10635; NAT2: P11245.

3.6.4. Splice site recognition and mRNA processing

Splice site effects were predicted using information theory-based analysis software (https://splice.cmh.edu), informing how the presence of a certain nucleotide can affect its recognition by the splice site machinery or affect recognition of the surrounding sequence or introduction of cryptic sites. As SNPs can also cause the introduction of pre-miRNA sites, this was included as part of the annotation in the novoSNP analysis procedure (see 3.3.).

3.6.5. Statistical analysis

3.6.5.1. Population differentiation

Genotype frequencies were calculated using Genepop (Raymond and Rousset, 1995) (http://genepop.curtin.edu.au) or Arlequin v3.11 (Excoffier et al., 2005) (http://cmpg.unibe.ch/software/arlequin3).

Probability value (p) was set at 0.05 for all statistical significance tests. $p<0.05$ means that less than 5% random differences are accepted and that at least 95% are true differences, indicating statistical significance.

Fisher's exact test (Fisher, 1922) was used to assess: (i) if there was significant deviation from HW proportions ($p<0.05$), the null hypothesis being that here is no difference between observed and expected genotype frequencies (ii) and if population differentiation (allele frequency differences among populations) was statistically significant ($p<0.05$), the null hypothesis being that the allelic distribution is identical across populations. The population differentiation p value was determined for frequency data of the re-

sequencing, HapMap SNP and commonly known alleles analyses, generating matrices of p values.

Re-sequencing data was assessed for genetic differentiation using Wright's **F statistic** (Wright, 1965), calculated according to (Weir and Cockerham, 1984). The fixation index Fst is a measure of variation among ethnic groups relative to the amount of variation expected under panmixia (where all populations are equally likely to mate with each other). Fst values range from 0 (no population subdivision, random mating occurrence, no genetic divergence within the population) to 1 (complete isolation or extreme division). For example, an Fst value of 0.02 shows that 2% of the total variation exists among populations, with the remainder, 98% of the variation occurring within populations. Negative Fst indicates no role of the respective loci in the genetic differentiation of populations. For each gene re-sequenced (*CYP2C9*, *CYP2C19*, *CYP2D6*, *NAT2*) pairwise Fst values of population pairs were used to generate an Fst matrix as estimates of genetic distance according to (Reynolds et al., 1983; Slatkin, 1995). Weighted Fst values obtained from the matrix give the average amount of genetic differentiation among populations.

3.6.5.2. Genetic and geographic distances

Assuming an **isolation by distance** model, the correlation between genetic and geographic distances was analysed. Genetic distance Fst values were converted to (Slatkin, 1993) measure of similarity and (Rousset, 1997) distance Fst/(1-Fst). A geographical distance matrix was then generated for population locations and converted to logarithmic (ln) distance. Reduced major axis (RMA) regression (Sokal and Rohlf, 1981) was used for plotting ln(dist) as x against Fst/1-Fst as y to determine the **correlation coefficient (r)** as the slope of the

equation (y=a+bx). **Mantel's test** (Mantel, 1967) was applied to assess whether the association between genetic similarity and geographic distance is statistically significant. Null hypothesis is that there is no relationship between geographic distance and genetic diversity, i.e. the slope of the best fit line from RMA is zero. If null hypothesis is rejected, conclusion would be that geographical distance affects genetic diversity. In order to assess the significance of this correlation (departure from slope=zero), random permutations are carried out, with correlation being calculated after each permutation. The reasoning is that if the null hypothesis of there being no relation between the two matrices is true, then permuting the rows and columns of the matrix should be equally likely to produce a larger or a smaller regression coefficient. The significance of the observed correlation is the proportion of such permutations that lead to a higher correlation coefficient, as represented by the **probability value (p)** at 95% confidence intervals. The **coefficient of determination (r^2)** is derived by regression from the x and y values of the equation (y=a+bx), minimising the distances between the squared vertical distances from the best fit line, and describes the proportion of genetic variation that can be explained by geographic distance.

Geographic coordinates were based on estimates of settlements or cities in recent settlement history as follows: Hausa (12:00:00N 8:31:00E); Yoruba (7:23:47N 3:55:00E); Igbo (5:23:00N 7:55:00E), Maasai (0:22:00S 36:05:00E), Luo (1:00:00S 33:00:00E); Shona, (17°51'50"S 31:01:47E); Kikuyu (0:09:00S 37:18:00E); Tanzanian Bantu (6:48:00S 39":17:00E); San (20:28:60S 27:49:0E); Venda, (24:0:0S 29:30:0E). Great circle distances measuring the shortest,

most direct route between two locations were determined to generate distance matrices.

3.6.5.3. Principal Component Analysis (PCA)

Principal Component Analysis (SIMCA-P+, www.umetrics.com) was applied to explore population relatedness using allele frequency variations of commonly known alleles. Only the first two principal components were determined, capturing the greatest variance among main world population clusters (Africans, Asians, Caucasians).

3.6.5.4. Phylogenetic analysis

Phylogeny Inference Package **PHYLIP** v3.65 (http://evolution.gs.washington.edu/phylip.html) was used for phylogenetic analysis of populations based on SNP frequency data. Pair-wise genetic distances were plotted with the ***gendist*** algorithm, while ***neighbour/UPGMA*** was used for plotting phylogenetic trees. The UPGMA method is according to Mary Kuhner and Jon Yamato and constructs a tree by successive clustering of lineages, setting branch lengths as the lineages join. The tree does not assume an evolutionary clock, so that it is in effect an un-rooted tree.

3.6.5.5. Analysis of Molecular Variance (AMOVA)

A hierarchical analysis of molecular variance (AMOVA), starting from the main world population clusters (Africans, Asians, Caucasians) and based on data sets from PCA and UPGMA, was performed using Arlequin v3.11 (Excoffier et al., 1992). AMOVA enables partition of genetic variance into components based on the following hierarchical subdivision: a=diversity among world population clusters; b=diversity among ethnic groups within population clusters; c=diversity within ethnic groups of all population clusters. Squared Euclidean distances

(distances between all pairs of populations based on a simple count of the number of differences in allele frequencies) are calculated, arranged into a matrix, and partitioned into submatrices corresponding to subdivisions of the population clusters. Squared distances from the Euclidean matrix are added up to give a total sum of squares. The same is repeated on the next level, to obtain for each population the *within* population cluster sum of squares. The *among* population cluster sum of squares is obtained from the squared distances amongst the population clusters. African population clusters were further analysed based on country, geographical region and ethno-linguistic family. Nested ANOVA (Analysis of Variance) was used to calculate how far individual squared distances are from the respective mean.

The variance components Va, Vb, Vc can be used to calculate a series of statistics called phi-statistics, which summarise the degree of differentiation between population divisions and are analogous to *F*-statistics. In this case, Fst values were determined to indicate the degree of differentiation amongst ethnic groups. The significance of the variance components and F-statistics was tested using 1,000 permutations with a significance value $p<0.05$.

3.7. Data records

3.7.1. Sample and genotype management

A sample and genotype recording system was designed specifically for this study, comprising a rich client interface (Borland Delphi 6 Enterprise Edition, Borland, Austin, TX) that accesses information stored in a Microsoft (MS) Access database (Microsoft, Redmond, WA).

Information is updated either via the rich client interface or through MS Access, subject to user authentication.

Samples are identified based on country of origin, ethnic group and a number (e.g. KNK76 is a sample from Kenya, of the Kikuyu ethnic group, number 76) and are annotated with the date of collection and/or DNA extraction as well as storage location.

RFLP genotyping data was entered manually into the MS Access database, while recording of data from re-sequencing, Taqman and Sequenom analyses was done through these technologies' inbuilt data management systems.

Sample and genotype information can be queried individually. Allele frequency calculations are generated upon request, exported into MS Excel (Microsoft, Redmond, WA) and used to build a catalogue.

3.7.2. Cataloguing and ranking of polymorphisms

A catalogue of African DME polymorphisms was established, detailing their impact on enzyme function and allele frequencies. A ranking system for SNPs was created, based on their frequency of occurrence in African populations as: <0.01=rare; 0.01-0.10=low; 0.10-0.30=medium; >0.30=high. If frequencies in different populations fell into several categories, hybrid rankings were made as rare/low, low/medium or medium/high.

4. RESULTS

Drug metabolising enzyme (DME) genes such as *CYP*s, *NAT*s, *GST*s and *FMO*s are highly polymorphic, and numerous SNPs and CNVs have been reported. The main theme of this study was to characterise DME gene variants in African populations by searching for unknown mutations as well as analyse HapMap SNPs and determine the prevalence of commonly known alleles.

Re-sequencing was used to detect novel variants of *CYP2C9*, *CYP2C19*, *CYP2D6* and *NAT2*. Polymorphisms were ascertained for each locus, and allele frequencies were calculated. The impact of SNPs on mRNA processing and protein function was estimated using bioinformatic applications.

HapMap SNPs were selected for major haplotypes of *CYP2B6*, *CYP2C9*, *CYP2C19* and *NAT2* gene regions. MALDI-TOF high-throughput multiplex genotyping (Sequenom) was employed for an analysis of three genes (*CYP2C9* was omitted due to difficulties in primer design) in four populations comparative to Yoruba from Ibadan (YRI), who were the only African representatives of the HapMap project.

The prevalence of several common alleles of *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *GSTM1*, *GSTT1 and NAT2*, was determined by RFLP, while *FMO3* mutations were studied using Taqman genotyping

Population differentiation and evolutionary relatedness were assessed based on DME allele frequencies, applying statistical methods such as Fischer's exact test, F-statistic, Mantel's test, Principal Component (PCA)/Phylogenetic (UPGMA) Analysis and AMOVA.

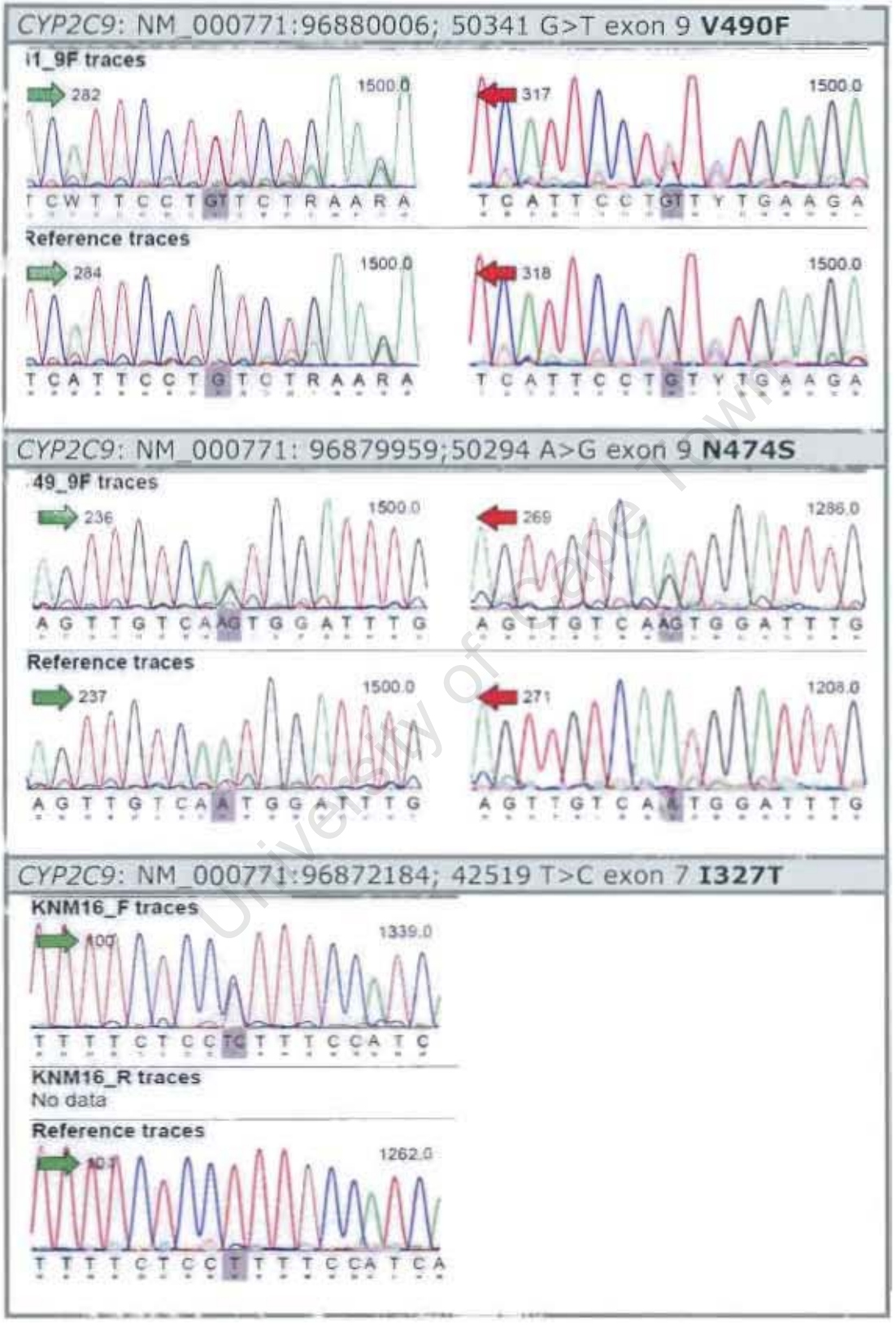
In summary, genotype data was generated for populations from three of the five main regions of Africa, representing four major ethno-linguistic classes (Table 1). Based on this diverse sample collection, a catalogue of African DME polymorphisms was created, marking a first step towards the development of a Pharmacogenetics Database of African Populations, and a pharmacodiagnostic kit was proposed as a contribution to the future determination of ADMET profiles in Africans.

4.1. Re-sequencing analysis of DME genes

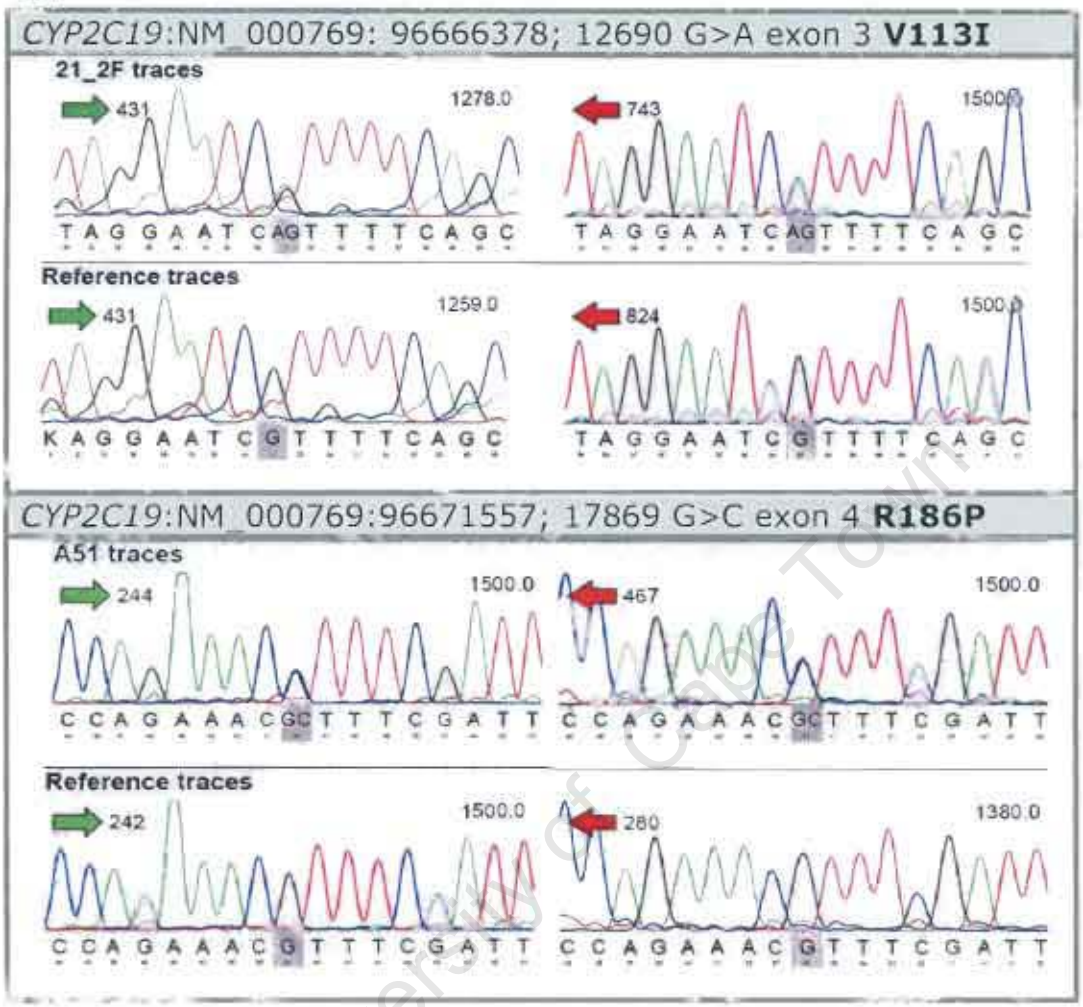
Drug metabolising enzyme genes *CYP2C9, CYP2C19, CYP2D6* and *NAT2* were re-sequenced in selected populations (see Tables 8-11). The sequencing covered all nine exons of each of the CYP genes and exon 2 of NAT2. Forward and reverse traces were produced for each SNP detected, except for *NAT2*, due to problems with the *NAT2* reverse sequencing primers. Hence only the *NAT2* forward sequencing reactions were successful and verified using a different set of forward primers.

Novel non-synonymous (Figure 3), synonymous and intronic SNPs were found in all four genes and their frequencies were determined (Tables 8-11). The potential impact of these polymorphisms on gene expression and/or enzyme function was assessed using bioinformatic applications.

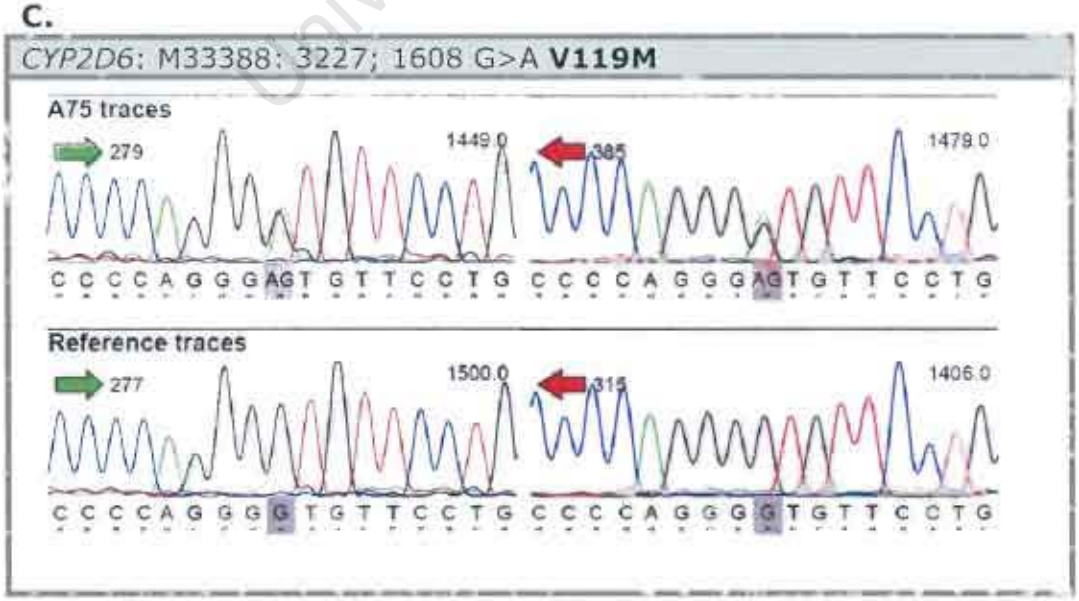
A.

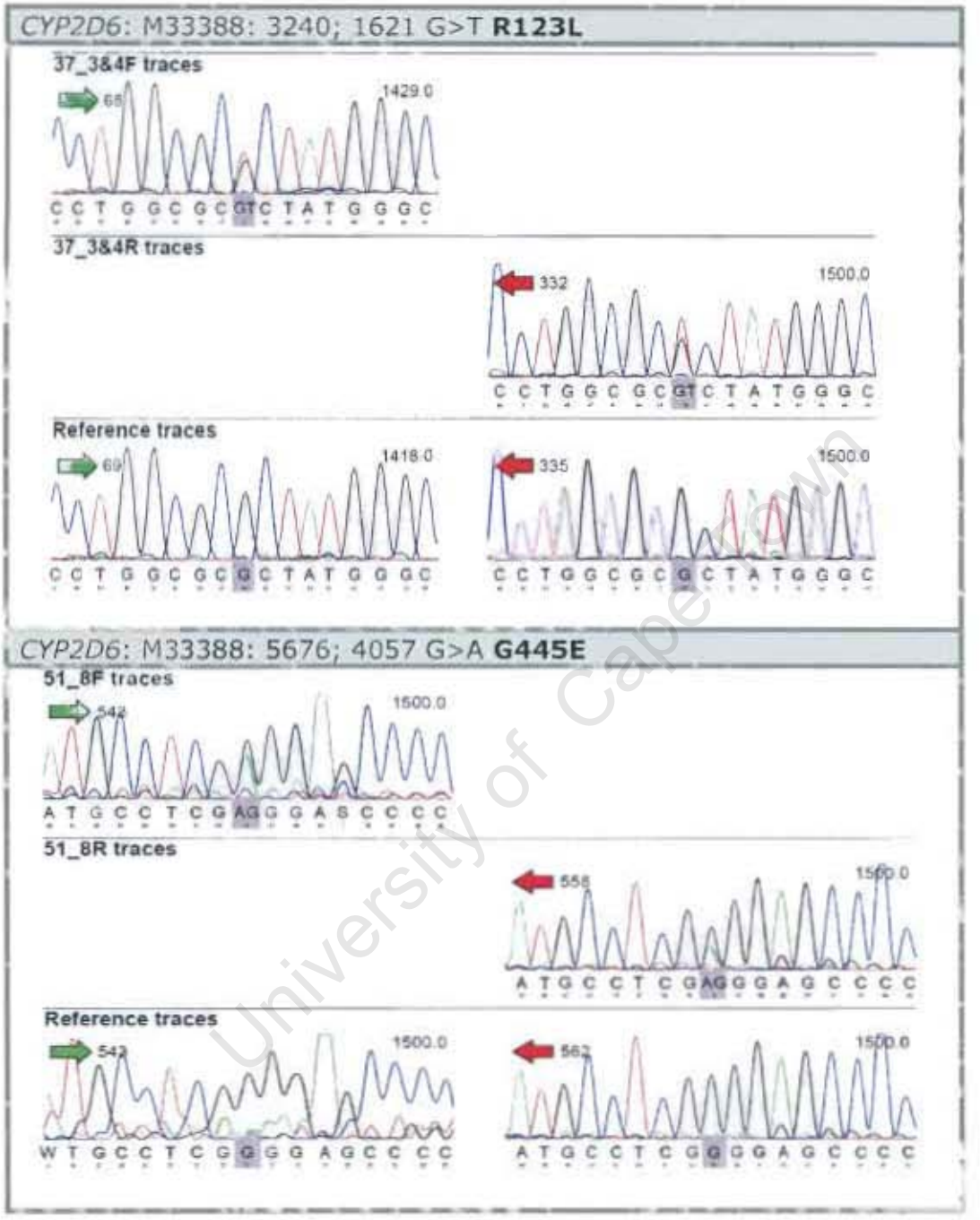


B.



C.





D.

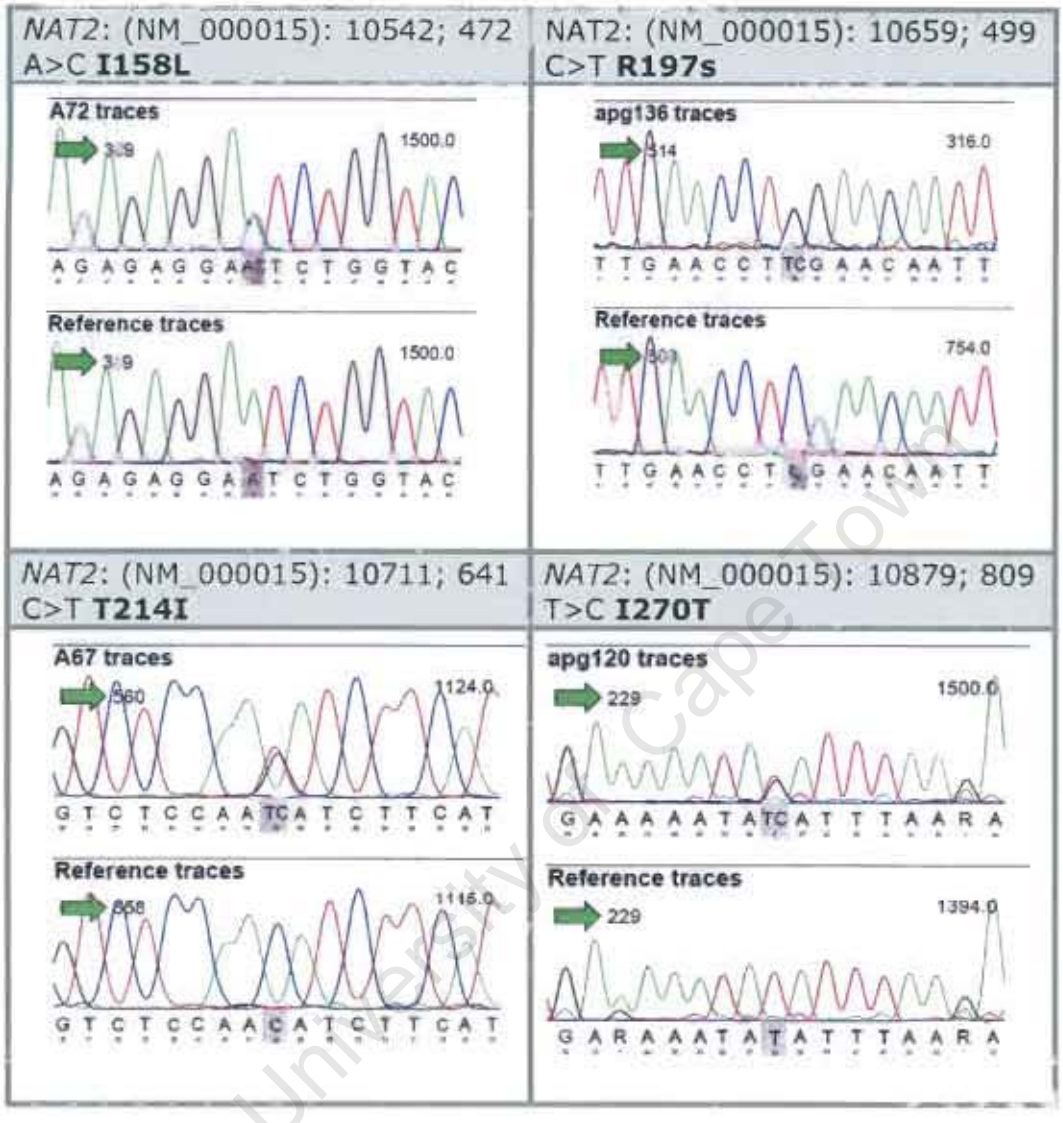


Figure 3: Sequence traces of novel non-synonymous SNPs **A.** *CYP2C9*; **B.** *CYP2C19*; **C.** *CYP2D6*; **D.** *NAT2*. Only forward sequence traces are available for *NAT2*.

Table 8: *CYP2C9* SNP frequencies

NC_000010.9 pos.	cDNA pos.	SNP	mRNA feature	effect	dbSNP	Hausa (13)	Luo (12)	Maasai (11)	San (13)	Shona (23)	Venda (9)	TZB (12)	Total (93)	p-value†	Fst value
96829291	-375	T>C	5'utr		rs9332103	0.04	nd	0.00	0.00	0.00	nd	nd	0.01	0.7064	-0.0055
96829916	251	T>C	intron		rs9332104	0.08	0.10	0.27	0.12	0.22	0.17	0.11	0.16	0.51248	0.0027
96833076	3411	T>C	intron		rs9332120	0.00	0.10	0.00	0.04	0.13	0.17	0.14	0.08	0.36011	0.0093
96833152	3487	A>G	intron		rs12769205	0.00	0.04	0.00	0.17	0.11	0.06	0.09	0.07	0.33028	0.0108
96833165	3499	T>A	intron		rs9332121	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	1	-0.0147
96838628	8963	T>C	intron		nrs	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.01	1	-0.0233
96838697	9032	G>C	intron		nrs	0.00	0.10	0.14	0.15	0.15	0.13	0.14	0.12	0.53058	-0.0122
96838734	9069	G>A	intron		novel	0.00	0.05	0.05	0.00	0.00	0.00	0.05	0.02	0.29747	-0.0044
96839116	9451	T>C	intron		rs17443251	0.00	0.05	0.00	0.00	0.00	0.00	0.06	0.01	0.4172	-0.0101
96839976	10311	A>G	intron		rs9332129	0.00	0.17	0.15	0.15	0.14	0.13	0.19	0.13	0.40538	-0.0114
96840012	10347	T>C	intron		novel	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	1	-0.0294
96840200	10535	A>G	exon 5	H251R (*9)	rs2256871	0.18	0.17	0.05	0.15	0.11	0.06	0.00	0.11	0.4758	0.0011
96840266	10601	wt> delA	exon 5	K273 fs (*6)	nrs	0.04	0.00	0.00	0.00	0.00	0.07	0.00	0.01	0.7091	-0.0055
96863014	33349	A>G	intron		rs9332172	0.17	0.23	0.23	0.15	0.28	0.25	0.50	0.26	0.23044	0.0209
96863323	33658	A>G	intron		rs9332174	0.13	0.14	0.27	0.12	0.20	0.19	0.10	0.17	0.77141	-0.0137
96872080	42415	C>T	intron		novel	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.33781	0.0044
96872134	42469	T>C	intron		rs9332197	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.69585	-0.0032
96872184	42519	T>C	exon 7	I327T	novel	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1	-0.0135
96872284	42619	G>C	exon 7	D360E (*5)	rs28371686	0.00	0.00	0.00	0.00	0.02	0.06	0.00	0.01	0.61909	-0.0069
96877210	47545	A>T	intron		rs9332230	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.66873	0.0015
96877258	47593	T>C	intron		rs9332232	0.07	0.14	0.05	0.04	0.18	0.17	0.05	0.11	0.43296	0.0023
96877304	47639	C>T	intron		rs2298037	0.00	0.00	0.00	0.00	0.03	0.06	0.00	0.01	0.54344	-0.0102
96879721	50056	A>T	intron		rs1934969	0.67	0.50	0.50	0.40	0.24	0.13	0.32	0.38	0.00696	0.0843
96879790	50125	C>T	intron		novel	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.67105	0.0015

Table 8: *CYP2C9* SNP frequencies *(continued)*

NC_000010.9 pos.	cDNA pos.	SNP	mRNA feature	effect	dbSNP	Hausa (13)	Luo (12)	Maasai (11)	San (13)	Shona (23)	Venda (9)	TZB (12)	Total (93)	p-value†	Fst value
96879861	50196	C>T	exon 9	A441A	rs2017319	0.04	0.14	0.05	0.04	0.20	0.17	0.05	0.10	0.18938	0.0212
96879959	50294	A>G	exon 9	N474S	novel	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.01	0.67189	0.0029
96879963	50298	A>T	exon 9	G475G	rs1057911	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.67435	0.0029
96880006	50341	G>T	exon 9	V490F	novel	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.66654	0.0015
96880078	50413	C>T	3'utr		rs9332240	0.00	0.05	0.00	0.00	0.03	0.00	0.00	0.01	1	-0.0228
96880099	50434	C>T	3'utr		rs9332241	0.08	0.05	0.00	0.00	0.02	0.13	0.14	0.05	0.12201	0.0232
96880166	50501	C>T	3'utr		rs9332243	0.00	0.05	0.00	0.00	0.03	0.00	0.00	0.01	1	-0.0228

†Overall population differentiation exact p-value=0.94042; Weighted Fst=0.0116; pos=position in sequence; cDNA pos=relative to A of ATG start codon; TZB=Tanzanian Bantu; wt=wild type; del=deletion; utr=untranslated region; fs=frameshift; (*)=described alleles carrying that particular mutation; nrs=rs number not yet assigned; nd=not determined; Number of individual samples studied per population is indicated; †Probability values based on Fischer's exact test for population differentiation; Fst value and weighted Fst is as explained in Materials and Methods.

Table 9: *CYP2C19* SNP frequencies

NC_000010.9 pos.	cDNA pos.	SNP	mRNA feature	effect	dbSNP	Hausa (20)	Yoruba (20)	Igbo (20)	Luo (30)	Maasai (13)	Shona (15)	Venda (9)	TZB (10)	Total (137)	p-value†	Fst value
96653591	-97	T>C	5'utr		rs4986894	0.13	0.18	0.33	0.07	0.08	0.20	0.11	0.15	0.15	0.05292	0.0279
96653743	55	A>C	exon 1	I19L (*15)	rs17882687	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.02	0.27245	0.0126
96653787	99	T>C	exon 1	P33P	rs17885098	0.05	0.08	0.15	0.18	0.15	0.17	0.17	0.30	0.15	0.23176	0.0081
96653871	183	T>C	intron		rs17882201	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	<0.01	0.06795	0.0318
96653876	188	G>A	intron		rs17881883	0.00	0.00	0.03	0.00	0.00	0.07	0.11	0.05	0.02	0.01753	0.0311
96653919	231	A>C	intron		novel	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	<0.01	1	-0.0157
96665810	12122	A>G	intron		rs7916649	0.50	0.37	0.29	0.58	0.50	0.25	0.28	0.33	0.41	0.05484	0.0313
96665994	12306	G>A	intron		rs17878649	0.00	0.00	0.03	0.08	0.04	0.10	0.06	0.10	0.05	0.19819	0.0079
96666148	12460	G>C	exon 2	E92D	rs17878459	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	<0.01	0.34849	0.0043
96666295	12607	wt>insC	intron		novel	0.00	0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.01	0.61739	-0.0023
96666325	12637	C>T	intron	splice site	novel	0.03	0.10	0.00	0.00	0.04	0.00	0.06	0.00	0.03	0.04129	0.0293
96666350	12662	A>G	intron	splice site	rs12769205	0.16	0.20	0.33	0.09	0.12	0.27	0.22	0.20	0.18	0.10753	0.0161
96666378	12690	G>A	exon 3	V113I	novel	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	<0.01	0.06371	0.0304
96666472	12784	G>A	exon 3	R144H (*9)	rs17884712	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	<0.01	0.06946	0.0312
96671557	17869	G>T	exon 4	R186P	novel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	<0.01	0.12223	0.0385
96671636	17948	G>A	exon 4	W212s (*3)	rs4986893	0.00	0.00	0.00	0	0.04	0.00	0.00	0.00	<0.01	0.22266	0.0131
96671895	18207	G>A	intron		novel	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.06	<0.01	0.37633	0.0065
96671917	18229	T>A	intron		rs17884938	0.06	0.03	0.05	0.07	0.00	0.10	0.00	0.00	0.05	0.70842	-0.0036
96671942	18254	T>C	intron		novel	0.03	0.03	0.00	0	0.00	0.00	0.00	0.00	0.01	0.90531	-0.0094
96672506	18818	T>C	intron	pre-miRNA	novel	0.00	0.03	0.00	0	0.00	0.00	0.00	0.00	<0.01	0.78515	-0.0057
96672599	18911	A>G	intron		rs7088784	0.05	0.08	0.15	0.14	0.15	0.17	0.17	0.25	0.13	0.44315	-0.0004
96672764	19076	T>C	intron	splice site	novel	0.00	0.00	0.00	0	0.00	0.03	0.00	0.00	<0.01	0.35105	0.0048
96672842	19154	G>A	exon 5	P227P (*2)	rs4244285	0.13	0.15	0.325	0.07	0.08	0.23	0.17	0.15	0.15	0.04419	0.0308
96673020	19332	G>A	intron	pre-miRNA	novel	0.00	0.00	0.00	0	0.08	0.00	0.00	0.00	<0.01	0.02948	0.0528

Table 9: *CYP2C19* SNP frequencies *(continued)*

NC_000010.9 pos.	cDNA pos.	SNP	mRNA feature	effect	dbSNP	Hausa (20)	Yoruba (20)	Igbo (20)	Luo (30)	Maasai (13)	Shona (15)	Venda (9)	TZB (10)	Total (137)	p-value†	Fst value
96711141	57453	G>C	intron		novel	nd	nd	0.10	0.04	0.00	0.13	0.00	0.00	0.04	0.36289	-0.009
96711200	57512	A>G	intron		novel	nd	nd	0.08	0.03	0.00	0.00	0.00	0.05	0.02	0.43404	-0.0062
96711255	57567	A>T	intron		novel	nd	nd	0.08	0.03	0.00	0.07	0.11	0.10	0.05	0.34088	-0.0057
96711263	57575	T>C	intron		novel	nd	nd	0.08	0.03	0.00	0.03	0.00	0.00	0.02	0.62689	-0.0111
96711325	57637	wt> delG	intron		novel	nd	nd	0.08	0.07	0.04	0.10	0.00	0.10	0.07	0.79314	-0.019
96711366	57678	T>G	intron		rs28399511	nd	nd	0.00	0	0.04	0.00	0.00	0.00	<0.01	0.4588	0.0057
96711428	57740	G>C	intron		rs4417205	nd	nd	0.08	0.09	0.13	0.23	0.22	0.20	0.14	0.36256	0.0044
96711677	57989	G>C	intron		novel	nd	nd	0.00	0.02	0.00	0.10	0.00	0.00	0.02	0.23719	0.0336
96733848	80160	C>T	exon 7	V330V	rs3758580	0.12	0.18	0.25	0.07	0.08	0.17	0.17	0.15	0.13	0.45878	-0.0032
96733849	80161	G>A	exon 7	V331I	rs3758581	0.03	0.03	0.00	0	0.04	0.00	0.00	0.00	0.01	0.59565	-0.0114
96734317	80629	T>A	intron		novel	0.03	0.05	0.00	0	0.05	0.00	0.00	0.00	0.01	0.46287	-0.0105
96740794	87106	T>C	intron		rs4917623	0.38	0.13	0.21	0.31	0.23	0.03	0.06	0.15	0.22	0.00394	0.0652
96740978	87290	T>C	exon 8	R410C (*13)	rs17879685	0	0.00	0.04	0.03	0.00	0.03	0.00	0.00	0.02	0.68325	-0.0083
96741001	87313	A>C	exon 8		rs17886522	0.03	0.05	0.08	0.07	0.04	0.10	0.00	0.10	0.06	0.7812	-0.012
96741110	87422	A>G	intron		novel	0.03	0.00	0.00	0	0.08	0.00	0.00	0.00	0.02	0.11062	0.0264
96741163	87475	G>C	intron		rs17880188	0.08	0.13	0.00	0.07	0.04	0.11	0.06	0.00	0.07	0.56199	-0.0004
96741210	87522	C>T	intron		rs17885567	0.05	0.08	0.21	0.10	0.08	0.13	0.11	0.10	0.1	0.67955	-0.009
96743266	89578	T>A	intron		rs12779363	0.00	0.00	0.06	0.04	0.00	0.03	0.00	0.00	0.01	0.6311	-0.0074
96743597	89909	C>T	intron		rs12268020	0.22	0.31	0.17	0.09	0.17	0.20	0.17	0.20	0.19	0.40663	-0.0005
96743699	90011	A>G	intron	splice site	rs4451645	0.13	0.14	0.00	0.07	0.04	0.13	0.11	0.00	0.08	0.42572	-0.0001
96743897	90209	A>C	exon 9	X491C; 26 extra aa (*12)	nrs	0.00	0.00	0.00	0.04	0.00	0.03	0.00	0.00	0.01	0.79593	-0.0066
96743989	90301	C>T	3'utr		novel	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	<0.01	0.00	0.00
96743990	90302	C>T	3'utr		novel	0.00	0.00	0.00	0.036	0.00	0.033	0.00	0.00	0.01	0.79123	-0.0062
96744221	90533	C>T	3'utr		novel	0.00	0.00	0.05	0.038	0.00	0.042	0.00	0.00	0.02	0.62085	-0.0072

†Overall population differentiation exact p-value=0.0097; Weighted Fst=0.0133; pos=position in sequence; cDNA pos=relative to A of ATG start codon; TZB=Tanzanian Bantu; wt=wild type; del=deletion; ins=insertion; utr=untranslated region; premiRNA=introduction of a premiRNA sequence; X=stop codon; aa=amino acid; (*)=described alleles carrying that particular mutation: nrs=rs number not yet assigned; nd=not determined; Number of individual samples studied per population is indicated; †Probability values based on Fischer's exact test for population differentiation; Fst value and weighted Fst is as explained in Materials and Methods.

Table 10: *CYP2D6* SNP frequencies

M33388 pos.	cDNA pos.	SNP	mRNA feature	effect	db SNP	Hausa (20)	Yoruba (20)	Igbo (20)	Luo (29)	Maasai (13)	Shona (15)	Venda (9)	TZA (10)	Total (136)	p-value†	Fst value
1444	-175	G>A	5'utr		rs1080993	0.05	0.12	0.31	0.22	0.05	0.11	0.06	0.30	0.17	0.07071	0.0238
1469	-150	C>T	5'utr		nrs	0.09	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02615	0.0414
1534	-85	T>C	5'utr		nrs	0.04	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.18896	0.0038
1577	-42	wt>insG	5'utr		rs28371695	0.19	0.35	0.13	0.21	0.05	0.20	0.13	0.10	0.19	0.09774	0.0183
1696	77	G>A	exon 1	R26H (*43)	rs28371696	0.04	0.00	0.03	0.00	0.00	0.03	0.00	0.05	0.02	0.52983	-0.0115
1701	82	C>T	exon 1	R28C (*22)	nrs	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	<0.01	0.60957	-0.0014
1719	100	C>T	exon 1	P34S (*10)	rs1065852	0.15	0.12	0.10	0.09	0.05	0.00	0.19	0.10	0.10	0.4074	-0.0179
1833	214	G>C	intron		rs1080995	0.50	0.41	0.27	0.36	0.45	0.57	0.17	0.43	0.38	0.47912	-0.0112
1840	221	C>A	intron		rs1080996	0.50	0.41	0.27	0.38	0.45	0.64	0.30	0.43	0.40	0.39003	-0.0013
1842	223	C>G	intron		rs1080997	0.50	0.41	0.27	0.36	0.45	0.56	0.20	0.38	0.38	0.47366	-0.0097
1846	227	T>C	intron		rs1080998	0.50	0.41	0.27	0.36	0.45	0.50	0.25	0.50	0.38	0.71714	-0.0291
1851	232	G>C	intron		rs1080999	0.50	0.41	0.30	0.36	0.50	0.70	0.25	0.75	0.42	0.34064	-0.0053
1852	233	A>C	intron		rs1080999	0.50	0.41	0.27	0.38	0.45	0.67	0.25	0.50	0.40	0.371	-0.0028
1864	245	A>G	intron		rs1081000	0.50	0.41	0.27	0.31	0.45	0.70	0.25	0.67	0.39	0.24599	0.0115
1929	310	G>T	intron		rs28371699	0.00	0.27	0.07	0.25	0.31	0.25	0.00	nd	0.18	0.08999	-0.0134
2273	654	C>T	intron		novel	nd	0.00	0.00	0.07	nd	0.08	0.07	0.07	0.07	1	-0.0454
2365	746	C>G	intron		nrs	nd	nd	nd	0.36	nd	0.40	0.31	0.50	0.40	0.96078	-0.0416
2462	843	T>G	intron		rs28371702	0.14	0.40	0.20	0.33	0.38	0.33	0.13	0.20	0.29	0.20191	-0.0011
2625	1006	C>T	exon 2	R101R	novel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.78433	-0.0089
2642	1023	C>T	exon 2	T107I (*17)	rs28371706	0.20	0.20	0.22	0.22	0.17	0.20	0.19	0.15	0.20	0.99623	-0.0377
2658	1039	C>T	exon 2	F112F	rs1081003	0.00	0.13	0.13	0.04	0.00	0.00	0.13	0.05	0.06	0.16493	0.0101
2686	1067	T>G	intron	splice site	novel	0.13	0.13	0.19	0.04	0.08	0.07	0.19	0.10	0.11	0.4348	-0.0072
3227	1608	G>A	exon 3	V119M	novel	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	<0.01	1	-0.0105
3240	1621	G>T	exon 3	R123L	novel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	<0.01	0.49312	0
3278	1659	G>A	exon 3	V136M (*29)	rs1058164	0.11	0.10	0.28	0.24	0.04	0.17	0.06	0.25	0.17	0.11449	0.0143
3280	1661	G>C	exon 3	V136V	rs28371708	0.29	0.43	0.35	0.32	0.46	0.37	0.33	0.30	0.35	0.75969	-0.0225
3335	1716	G>A	exon 3	E155K (*45)	rs28371710	0.08	0.00	0.00	0.09	0.04	0.00	0.17	0.00	0.05	0.08641	0.0087
3465	1846	G>A	intron	182 splicing defect (*4)	nrs	0.03	0.08	0.08	0.04	0.04	0.00	0.00	0.05	0.04	0.72795	-0.0186
3483	1864	wt>delGT	exon 4	ins 3 aa	nrs	0.00	0.00	0.00	0.04	0.08	0.00	0.00	0.00	0.02	0.07725	0.0279
3485	1866	C>T	exon 4	N175N	nrs	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	<0.01	1	-0.0105
3488	1869	T>C	exon 4	G176G	nrs	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	<0.01	0.78389	-0.0016

Table 10: *CYP2D6* SNP frequencies *(continued)*

M33388 pos.	cDNA pos.	SNP	mRNA feature	effect	db SNP	Hausa (20)	Yoruba (20)	Igbo (20)	Luo (29)	Maasai (13)	Shona (15)	Venda (9)	TZA (10)	Total (136)	p-value†	Fst value
3617	1998	T>C	exon 4	F219F	novel	nd	nd	0.00	0.00	nd	0.03	0.00	0.05	0.02	1	-0.0441
4194	2575	C>A	exon 5	P267P	nrs	nd	nd	nd	0.05	nd	0.03	0.22	0.00	0.07	0.56492	-0.0182
4221	2602	G>T	exon 5	L276L	novel	nd	nd	nd	0.05	nd	0.00	0.06	0.00	0.02	0.71336	-0.0132
4280	2661	G>A	intron		nrs	nd	nd	nd	0.05	nd	0.03	0.11	0.05	0.06	0.84608	-0.0369
4379	2760	T>A	intron		novel	nd	nd	nd	0.00	nd	0.10	0.06	0	0.04	0.21983	0.0271
4469	2850	C>T	exon 6	R296C	nrs	nd	nd	nd	0.55	nd	0.63	0.44	0.65	0.58	0.76292	-0.0362
4607	2988	G>A	intron	splicing defect	nrs	nd	nd	nd	0.00	nd	0.03	0.00	0.00	0.01	0.57727	0
4802	3183	G>A	exon 7	V338M (*29)	nrs	0.13	0.10	0.29	0.20	0.04	0.17	0.06	0.13	0.16	0.12732	0.0171
4873	3254	T>C	exon 7	H361H	rs2743457	0.09	0.00	0.00	0.07	0.08	0.00	0.13	0.00	0.04	0.08139	0.0061
4880	3259	wt>insTG	exon 7	375 fs (*42)	nrs	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	<0.01	0.21036	0.0069
5003	3384	A>C	intron		nrs	0.30	0.45	0.34	0.28	0.42	0.37	0.25	0.38	0.65	0.63972	-0.0196
5016	3397	C>A	intron		novel	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	<0.01	0.4771	0.0037
5180	3561	G>C	intron		novel	0.00	0.00	0.00	0.02	0.00	0.00	0.06	0.06	0.01	0.38425	0.0127
5201	3582	A>G	intron		nrs	0.08	0.11	0.11	0.09	0.04	0.00	0.13	0.00	0.08	0.61433	-0.0223
5203	3584	G>A	intron		nrs	0.54	0.34	0.26	0.43	0.46	0.47	0.44	0.44	0.41	0.37247	-0.0013
5326	3707	G>A	intron		nrs	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.01	0.56256	-0.0041
5349	3721	wt>delGT	intron		nrs	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	<0.01	0.78736	-0.0035
5409	3790	C>T	intron	splice site	nrs	0.53	0.34	0.26	0.44	0.54	0.47	0.44	0.44	0.42	0.22202	0.0076
5472	3853	G>A	exon 8	E410K (*27)	nrs	0.00	0.00	0.00	0.06	0.04	0.00	0.00	0.06	0.02	0.3001	0.0066
5652	4033	C>T	intron	splice site	novel	0.00	0.00	0.00	0.02	0.00	0.00	0.06	0.06	0.01	0.25752	0.0156
5676	4057	G>A	exon 9	G445E	novel	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.01	0.65538	0.0068
5799	4180	G>C	exon 9	S486T	rs1135850	0.68	0.55	0.66	0.72	0.63	0.63	0.75	0.67	0.66	0.74529	-0.021
6013	4394	wt>delAG	3'utr		novel	0.00	0.00	0.00	0.00	0.00	0.10	0.06	0.00	0.02	0.00774	0.0541
6020	4401	C>T	3'utr		nrs	0.09	0.10	0.11	0.07	0.04	0.00	0.25	0.06	0.08	0.47987	-0.0099
6100	4481	G>A	3'utr		nrs	0.12	0.08	0.03	0.11	0.21	0.23	0.07	0.19	0.12	0.12074	0.0199
6275	4656	wt>delACA	3'utr		nrs	0.44	0.09	0.03	0.23	0.31	0.08	0.33	0.33	0.20	0.00132	0.0978
6341	4722	T>G	3'utr		nrs	0.63	0.57	0.73	0.57	nd	nd	nd	0.25	0.58	0.08791	0.0199

†Overall population differentiation exact p-value <<0.00001; Weighted Fst=-0.0048; pos=position in sequence; cDNA pos=relative to A of ATG start codon; TZB=Tanzanian Bantu; wt=wild type; del=deletion; ins=insertion; utr=untranslated region; s=stop codon; (*)=described alleles carrying that particular mutation; fs=frame shift; aa=amino acid; nrs=rs number not yet assigned; nd=not determined; Number of individual samples studied per population is indicated; †Probability values based on Fischer's exact test for population differentiation; Fst value and weighted Fst is as explained in Materials and Methods.

Table 11: *NAT2* SNP frequencies

NC_000008.9 pos.	cDNA pos.	SNP	effect	db SNP	Hausa (20)	Yoruba (20)	Igbo (19)	Luo (16)	Maasai (12)	San (40)	Total (125)	p-value†	Fst value
8950	191	G>A	R64Q (*14)	rs1801279	0.03	0.08	0.13	0.20	0.08	0.09	0.1	0.24379	0.0068
9041	282	C>T	Y94Y	rs1041983	0.40	0.44	0.55	0.48	0.38	0.29	0.39	0.12467	0.0173
9100	341	T>C	I114T (*5)	rs1801280	0.33	0.14	0.34	0.27	0.50	0.20	0.27	0.02245	0.0422
9162	403	C>G	L135V	nrs	0.00	0.03	0.00	0.03	0.00	0.00	<0.01	0.23886	0.0033
9231	472	A>C	I158L	novel	0.00	0.00	0.00	0.03	0.00	0.00	<0.01	0.21833	0.0119
9240	481	C>T	L161L	rs1799929	0.25	0.14	0.34	0.27	0.46	0.14	0.24	0.00753	0.0535
9348	589	C>T	R197X	novel	0.00	0.00	0.00	0.00	0.00	0.01	<0.01	1	-0.015
9349	590	G>A	R197Q (*6)	rs1799930	0.32	0.33	0.29	0.30	0.25	0.20	0.27	0.59091	-0.0077
9400	641	C>T	T214I	novel	0.00	0.00	0.00	0.03	0.00	0.00	<0.01	0.22588	0.0094
9442	683	C>T	P228L	nrs	0.00	0.00	0.00	0.03	0.00	0.00	<0.01	0.22841	0.0094
9525	766	A>G	K256E	nrs	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.89704	-0.0037
9562	803	A>G	K268R	rs1208	0.37	0.39	0.40	0.44	0.54	0.43	0.42	0.82844	-0.0126
9568	809	T>C	I270T	novel	0.00	0.00	0.00	0.00	0.00	0.13	0.04	0.00244	0.0875
9597	838	G>A	V280M	nrs	0.06	0.03	0.05	0.03	0.00	0.00	0.02	0.14861	0.0027
9616	857	G>A	G286E (*7)	rs1799931	0.03	0.03	0.03	0.03	0.04	0.01	0.02	0.88729	-0.0214

Overall population differentiation exact p-value =0.00469; Weighted Fst=0.0169; pos=position in sequence; cDNA pos=relative to A of ATG start codon; X=stop codon; (*)=described alleles carrying that particular mutation; nrs=rs number not yet assigned; Number of individual samples studied per population is indicated; †Probability values based on Fischer's exact test for population differentiation; Fst value and weighted Fst is as explained in Materials and Methods.

4.1.1. Characterisation of non-synonymous SNPs

In *CYP2C9* (Table 8), three out of six non-synonymous SNPs were novel: 42519 T>C (I327T), 50294 A>G (N474S) and 50341 G>T (V490F), of which I327T and V490F are predicted to have a functional effect (Table 12). However, further inference of these amino acid changes with crystal structure information (Williams et al., 2003b) and sequence alignments (Gotoh, 1992; Mestres, 2005) indicates that they may not influence substrate recognition and binding.

The two novel non-synonymous SNPs discovered in *CYP2C19* (Table 9), 12690 G>A (V113I) in exon 3 and 17869 G>C (R186P) in exon 4, seem to cause very different effects on enzyme function, according to the physicochemical character of their amino acid changes (Table 12). Whereas the effect of V113I may be negligible, the change from the basic arginine to proline at position 186 was predicted as functionally damaging (PSIC score=3.159).

Three novel non-synonymous SNPs were found in *CYP2D6*: 1608 G>A (V119M), 1621 G>T (R123L) and 4057 G>A (G445E) (Table 10), of which V119M and R123L were predicted to have no effect on enzyme function (Table 12), although they are located in the substrate recognition site SRS1. The G445E substitution may be functionally important (PSIC score=3.063), owing to its close proximity to the 443 site, which is critical for heme ligand binding in this enzyme, according to the crystal structure (Rowland et al., 2006a). Consistent with other African data (Masimirembwa et al., 1996; Wennerholm et al., 2001), the most common non-synonymous SNPs that contribute to variation in drug response were 2850 C>T; 4180 G>C (R296C; S486T), 1023 C>T (T107I) and 1659 G>A; 3183 G>A (V136M; V338M),

corresponding to the commonly known alleles *CYP2D6*2*, *CYP2D6*17* and *CYP2D6*29*, respectively.

Four novel amino acid-changing SNPs were detected in *NAT2*: 472 A>C (I158L), 589 C>T (R197X), 641 C>T (T214I) and 809 T>C (I270T) (Table 11). The 641 C>T (T214I) variant was predicted to have an effect on enzyme function (Table 12), because the amino acid at this position would be involved in coenzyme A ligand binding as part of the acetylation process. The 589 C>T (R197X) mutation results in a stop codon being introduced and hence no protein expressed. The most common alleles of *NAT2* in this study were *NAT2*5* (341 T>C, I114T) and *NAT2*6* (590 G>A, R197Q) (Table 11), which contribute largely to the slow acetylator phenotype in African populations.

4.1.2. Determination of splice variants

In addition to non-synonymous SNPs, numerous novel synonymous SNPs, SNPs in introns and at splice site junctions, were identified (Tables 8-11). Novel SNPs in splice site junctions were investigated, but none of those were located within the most critical 0 to -2 positions of the acceptor sites or the –2 to +4 positions of the donor sites. Known SNPs affecting splicing include *CYP2D6*4* 1846 G>A at position 0 of exon 4 and the synonymous SNP *CYP2C19*2* 19154 G>A located 40 base pairs from the start of exon 5, resulting in a cryptic splice site.

4.1.3. Haplotype determination

Novel SNPs for *CYP2C9*, *CYP2C19* and *CYP2D6* were grouped and assigned to haplotypes or groups of other known mutations if possible

(Table 12). New allele names are based on non-synonymous SNPs, or those well validated to affect enzyme expression and function (e.g. mRNA processing or in promoter regions).

CYP2C9 42519 T>C (I327T) and 50341 G>T (V490F) were assigned new allele names as *CYP2C9*31* and *CYP2C9*32*, respectively. For *CYP2C19*, the 17869 G>C/80161 A>G (R186P/I331V) combination was assigned the new allele *CYP2C19*22*. It was not possible to assign new haplotypes/alleles for *CYP2C9* 50294 A>G (N474S) and *CYP2C19* 12690 G>A (V113I), because their linkage with other alleles such as *CYP2C9*9* 10535 A>G (H251R) and *CYP2C19*2* 19154 G>A, respectively, could not be excluded. Some synonymous SNPs, as well as non-coding SNPs, were grouped to known mutations such as *CYP2C9*9*, *CYP2C19*12* and *CYP2C19*13* (Table 12). It appears that the novel *CYP2D6* 1608 G>A (V119M) is found on the known *CYP2D6*29* allele, which is defined by 1659 G>A (V136M) and 3183 G>A (V338M). This haplotype group was therefore assigned the new name *CYP2D6*70*. New alleles were not assigned for *CYP2D6* 1621 G>T (R123L) and 4057 G>A (G445E), pending an expanded haplotype analysis.

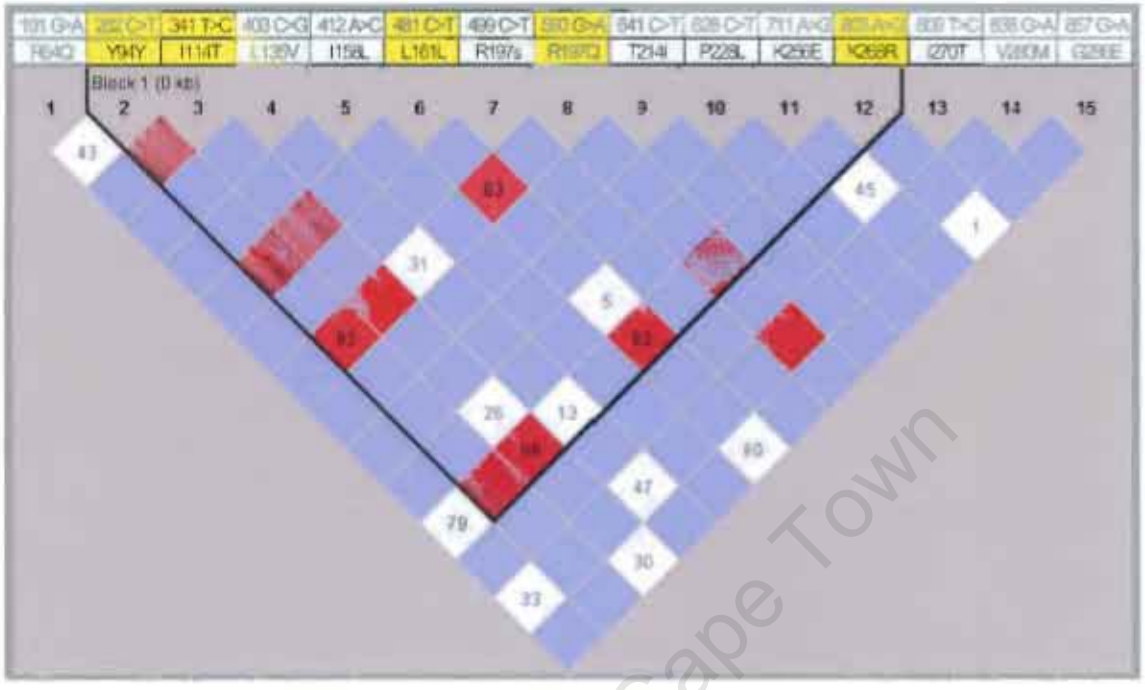
The novel *NAT2* SNPs did not appear to be linked to any other SNPs. Haploview-determined tagSNPs for *NAT2* were used to determine the major haplotype frequencies (Figure 4). The most common sub-haplotypes were *NAT2*6A* and *NAT2*5B*, which affect enzyme function, followed by the wild type *NAT2*4* and *NAT2*12A*, which do not impair acetylation.

Table 12: Grouping of novel SNPs and functional effect prediction

Gene	SNP grouping	cDNA pos.	Amino acid change	Functional effect prediction (PSIC score)
CYP2C9	**42519 T>C (*31)**	**980**	**I327T**	**Possible functional damage (2.761)**
	50294 A>G	**1421**	**N474S**	**No functional damage (0.162)**
	47545 A>T 50298 A>T **50341 G>T (*32)**	**1468**	**V490F**	**Possible functional damage (1.806)**
	*10535 A>G (*9)* 50196 C>T	*752* 1323	*H251R* A441A	*Possible functional damage (2.239)*
CYP2C19	12122 G>A **12690 G>A**	**337**	**V113I**	**No functional damage (0.198)**
	57453 G>C ***90533 C>T*** ***57575 T>C*** 87290 C>T (*13)	1228	R410C	
	*90209 A>C (*12)* ***90302 C>T***	*1473*	*X491C*	*26 extra amino acids*
	17869 G>C (*22) *80161 G>A*	**557** *991*	**R186P** *I331V*	**Possible functional damage (3.159)**
CYP2D6	-175 G>A 310 G>T 843 T>G **1608 G>A (*70)** *1659 G>A* 1661 G>C *3183 G>A* 3384 A>C *4180 G>C* 4722 T>G		 **V119M** *V136M* *V338M* *S486T*	 **No functional damage (0.054)** *Functional damage-reduced enzyme activity* *Functional damage-reduced enzyme activity* *No functional damage (0.267)*
	214 G>C 223 C>G 227 T>C 843 T>G **1621 G>T** 1661 G>C *2850 C>T* 3384 A>C 3584 G>A 3790 C>T *4180 G>C*		 **R123L** *R296C* *S486T*	 **No functional damage (1.236)** *No functional damage (0.254)* *No functional damage (0.267)*
	843 T>G 1661 G>C *2850 C>T* 3384 A>C **4057 G>A** *4180 G>C*		 *R296C* **G445E** *S486T*	 *No functional damage (0.254)* **Possible functional damage, contact with functional site (3.063)** *No functional damage (0.267)*
NAT2	**10542 A>C**	**472**	**I158L**	**No functional damage (0.615)**
	10659 C>T	**589**	**R197X**	**No protein expressed**
	10711 C>T	**641**	**T214I**	**Possible functional damage, involved in ligand binding (1.257)**
	10879 T>C	**809**	**I270T**	**No functional damage (0.526)**

PSIC=Position-Specific Independent Counts; cDNA pos=relative to A of ATG start codon; (*)=described alleles carrying that particular mutation; bold: novel non-synonymous SNPs; italic bold: novel intronic SNPs; italics: known non-synonymous SNPs; SNP positions are according to reference sequences (Tables 8-11).

A



B

282 C>T	341 T>C	481 C>T	590 G>A	803 A>G	haplotype	phenotype	frequency
√			√		*6A	slow	0.25
	√	√		√	*5B	slow	0.23
					*4	fast	0.17
				√	*12A	fast	0.15
√					-	fast	0.14
	√			√	*5C	slow	0.04
			√		*6B	slow	0.01

Figure 4: *NAT2* haplotypes constructed from sequence and genotype data of the total population studied (n=127).
A: LD plot with the genomic positions indicated at the top. In yellow are the tagSNPs which define the major known haplotypes and are able to capture other SNPs within the same haplotype. The amino acid changes at the various positions are shown.
B: Haplotype frequencies. Haplotypes and phenotypes (acetylators) were assigned according to Hein et al., 2008).

4.2. Analysis of HapMap SNPs

As the Yoruba from Ibadan (YRI) were the only population representing Africans in the HapMap project, this study aimed at assessing the applicability to other Africans of Yoruba SNPs in major haplotype blocks of *CYP2B6*, *CYP2C9*, *CYP2C19* and *NAT2*. Four populations (Maasai, Hausa, San and Shona) were studied, representing eastern, western and southern Africa.

Selection criteria of tagSNPs were based on their significance in tagging major haplotypes. Using the Haploview Tagger tool (see 3.6.2), stringency was set at pair-wise regression of SNPs giving $r^2>0.8$. SNPs within some blocks were selected after setting haplotype frequencies at 0.01 and above. Throughout the course of the analysis, primer design and experimental conditions were checked for consistency. This resulted in the removal of several SNPs due to some regions being non-specific or including repeat sequences, causing difficulties in primer design. In conclusion, 30 SNPs were analysed successfully: 10 for *CYP2B6*, 9 for *CYP2C19*, and 11 for *NAT2* (Table 13) but none for *CYP2C9*. It should be noted that some of these SNPs were also analysed by other methods (re-sequencing, RFLP).

Table 13: Genotype frequencies of HapMap-based SNPs

Gene	SNP	Maasai (51)	Hausa (48)	San (40)	Shona (46)	YRI¶ (60)	p-value†	Fst value
CYP2B6	rs10426235	0	0.01	0.038	0.067	0.08	0.01628	0.0069
	rs12721649	0.112	0.198	0.2	0.222	0.26	0.18108	0.0224
	rs2279343	0.43	0.398	0.473	0.42	0.55	0.82435	0.0008
	rs2306606	0.36	0.489	0.4	0.4	0.46	0.31439	-0.0088
	rs3745274	0.598	0.67	0.606	0.692	0.46	0.52737	-0.0042
	rs3786547	0.36	0.435	0.363	0.389	0.34	0.71471	0.0003
	rs7259965	0.375	0.396	0.338	0.356	0.35	0.86949	0.0031
	rs8100458	0.245	0.219	0.237	0.089	0.12	0.01549	0.0019
	rs8192711	0.106	0.022	0.039	0.044	0.03	0.08023	-0.0011
	rs8192712	0.041	0	0.013	0	0.03	0.04217	0.0183
CYP2C19	rs10509676	0.511	0.462	0.5	0.5	0.14	0.93526	0.0013
	rs11597626	0.223	0.196	0.329	0.244	0.24	0.24689	0.0097
	rs12779363	0.204	0.188	0.213	0.222	0.28	0.94655	0.0295
	rs4244285	0.122	0.117	0.087	0.122	0.14	0.87744	0.0235
	rs4304697	0.1	0.094	0.025	0.111	0.08	0.12984	0.0107
	rs4388808	0.224	0.234	0.087	0.109	0.17	0.01035	0.0768
	rs4986893	0.04	0.011	0	0	0	0.07283	0
	rs4986894	0.11	0.096	0.075	0.109	0.14	0.86051	0.0022
	rs7088784	0.115	0.106	0.025	0.133	0.13	0.05484	0.0057
NAT2	rs1208	0.553	0.391	0.434	0.435	0.41	0.15493	0.0574
	rs1799929	0.468	0.25	0.145	0.256	0.19	**0**	0.0056
	rs1799930	0.276	0.287	0.188	0.244	0.18	0.43708	-0.0078
	rs1799931	0.039	0.042	0.013	0.022	0.05	0.664	-0.0079
	rs1801279	0.02	0.089	0.092	0.133	0.09	0.02252	0.0169
	rs1801280	0.5	0.293	0.205	0.283	0.28	**0.00017**	0.062
	rs4646246	0.092	0.277	0.154	0.322	0.32	**0.00039**	0.0526
	rs4646247	0.27	0.272	0.138	0.244	0.18	0.11437	0.0112
	rs721398	0.27	0.266	0.188	0.189	0.18	0.3691	-0.0088
	rs721399	0.21	0.389	0.438	0.489	0.54	**0.0001**	-0.0068
	rs7832071	0.54	0.337	0.25	0.433	0.42	**0.00041**	0.0547

†Overall population differentiation exact p-value=<<0.0001; Weighted Fst=0.0166 ¶SNP frequency data obtained from HapMap database; †Probability values based on Fischer's exact test for population differentiation; Fst value and weighted Fst is as explained in Materials and Methods; Bold=statistically significant allele frequency differences between populations.

4.3. Baseline prevalence of common alleles

Common variants of *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *FMO3 and NAT2* are well characterised with functional and resultant clinical associations extensively explored. Such SNPs were analysed, using RFLP or Taqman genotyping, to compare their baseline prevalence in Africans with literature data for Caucasian and Asian populations (Table 14).

For *GSTM1* and *GSTT1*, only homozygous deletion genotypes were determined. Using the PCR method, it was not possible to determine the heterozygous condition. However, it is well established that the homozygous deletion is responsible for adverse clinical characteristics.

*CYP2B6*6* is highly prevalent in the African populations of this study, with frequencies ranging from 0.34-0.42. This data is consistent with the prevalence of *CYP2B6*6* in other African populations, with frequencies of 0.49 and 0.47 reported in Ghanaians and African Americans, respectively. Overall, African figures were considerably higher than the frequencies of *CYP2B6*6* reported in Caucasian and Asian populations (Cho et al., 2004; Hiratsuka et al., 2002; Lang et al., 2004).

*CYP2C9*2* and *CYP2C9*3* were not found in the African populations of this study, in agreement with their absence or low frequencies in other Africans and African Americans. Although the prevalence of *CYP2C9*2* and *CYP2C9*3* is also low in Asians and Caucasians, they are the main alleles associated with reduced enzyme function in these populations.

Table 14: Frequencies of commonly known alleles in African populations from this study and in Caucasians and Asians from literature sources

Gene	Allele	Kikuyu	Luo	Maasai	Igbo	Yoruba	Hausa	San	Shona	Venda	TZB	Gha	Eth	AfA	Swe	Ger	Chi	Jap	Kor
CYP2B6	*6	**0.34**	**0.37**	**0.35**	**0.38**	**0.42**	**0.42**	**0.40**	**0.38**	**0.36**	**0.39**	0.49	*0.35*	0.47	*0.21*	*0.21*	0.21	0.16	0.15
CYP2C9¶	*2	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	*0.01*	*0.03*	*0.03*	0.12	0.12	0.00	0.00	0.00
	*3	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	*0.00*	*0.00*	*0.00*	0.07	0.07	0.05	0.05	0.06
CYP2C19	*2	**0.16**	**0.18**	**0.11**	**0.29**	**0.10**	**0.12**	**0.12**	**0.13**	0.21	**0.18**	*0.15*	0.14	0.25	0.17	0.18	0.37	0.35	0.21
	*3	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.00	**0.01**	*0.00*	0.02	0.00	0.00	0.00	0.08	0.11	0.12
CYP2D6	*2XN	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	0.02	*0.02*	*0.03*	0.02	0.29	0.01	0.01	0.02	0.01	0.01	0.00
	*3	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.00	**0.00**	0.00	0.00	*0.00*	0.03	0.02	0.00	0.00	0.00
	*4	**0.01**	**0.04**	**0.08**	**0.08**	**0.03**	**0.02**	**0.09**	**0.02**	0.03	**0.02**	0.07	0.04	0.07	0.23	0.20	0.01	0.01	0.02
	*5	*0.03*	*0.02*	*0.01*	*0.05*	*0.04*	*0.05*	**0.01**	0.04	0.05	0.04	0.06	0.03	0.06	0.05	0.02	0.06	0.03	0.06
	*9	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.00	**0.00**	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
	*10	*0.05*	**0.06**	**0.05**	**0.10**	**0.07**	**0.13**	*0.05*	**0.06**	**0.12**	**0.04**	0.03	0.09	0.04	0.01	0.02	0.51	0.43	0.51
	*17	**0.33**	**0.23**	**0.18**	**0.14**	**0.22**	**0.18**	**0.22**	**0.34**	0.24	**0.18**	0.28	0.09	0.15	0.00	0.00	0.00	0.00	0.00
	*29	**0.14**	**0.16**	**0.08**	**0.20**	**0.10**	**0.10**	**0.02**	**0.17**	**0.06**	**0.20**	*0.15*	*0.20*	0.05	0.00	0.00	0.00	0.00	0.00
FMO3¶	E158K	**0.49**	**0.50**	**0.42**	**0.44**	**0.52**	**0.44**	**0.33**	**0.50**	**0.48**	*0.50*	*0.48*	*0.48*	0.42	0.44	0.43	0.23	0.23	0.19
	V257M	**0.05**	**0.05**	**0.04**	**0.01**	**0.01**	**0.04**	**0.01**	**0.02**	**0.02**	*0.02*	*0.02*	*0.02*	0.07	0.07	0.07	0.20	0.15	*0.15*
	E308G	**0.01**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**	**0.01**	**0.02**	**0.02**	*0.01*	*0.01*	*0.01*	0.05	0.22	0.23	0.15	0.21	0.18
	L360P	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
	D132H	**0.01**	**0.02**	**0.03**	**0.08**	**0.07**	**0.05**	**0.02**	**0.04**	**0.03**	*0.03*	*0.03*	*0.03*	*0.03*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
GST	M1 del	**0.28**	**0.29**	**0.16**	**0.23**	**0.31**	**0.37**	**0.45**	**0.24**	0.23	**0.33**	0.39	*0.38*	0.28	0.51	0.51	0.58	0.44	0.53
	T1 del	**0.25**	**0.22**	**0.40**	**0.36**	**0.35**	**0.42**	*0.25*	**0.26**	0.20	**0.25**	*0.23*	*0.23*	0.24	0.20	0.21	0.53	0.44	0.60
NAT2	*5	*0.33*	**0.34**	**0.42**	**0.28**	**0.33**	**0.27**	**0.20**	**0.31**	0.39	**0.34**	*0.32*	*0.33*	0.30	0.51	0.46	0.06	0.02	0.03
	*6	*0.24*	**0.22**	**0.27**	**0.29**	**0.27**	**0.33**	**0.08**	**0.21**	0.22	**0.21**	*0.21*	*0.22*	0.22	0.28	0.27	0.31	0.19	0.19
	*7	*0.02*	**0.03**	**0.04**	**0.04**	**0.03**	**0.03**	**0.01**	**0.06**	0.05	**0.03**	*0.04*	*0.02*	0.02	0.02	0.04	0.16	0.10	0.11
	*14	*0.11*	**0.14**	**0.09**	**0.11**	**0.08**	**0.03**	**0.09**	**0.14**	0.11	**0.13**	*0.12*	*0.09*	0.09	0.00	0.00	0.00	0.00	0.00

TZB=Tanzanian Bantu; Gha=Ghanaian, Eth=Ethiopian; AfA=African American; Swe=Swede; Ger=German, Chi=Chinese, Jap=Japanese; Kor=Korean; Black and bold=frequencies obtained from RFLP genotyping; Blue and bold=frequencies obtained from re-sequencing genotype data; Regular=frequencies obtained from literature; *2XN=multiplication alleles of CYP2D6; Italics=in case of lacking frequency data in literature, frequencies were estimated based on average of closely related ethno-linguistic groups belonging to the same family or class, as complete datasets are required for population clustering and phylogenetic analysis; ¶data were not included in PCA analysis.

*CYP2C19*2* frequencies ranged from 0.10 to 0.29 and were highest in Igbos. This data falls within the range (0.13 to 0.25) reported for other Africans and African Americans (Xie et al., 1999a). *CYP2C19*3* was detected at low frequencies in Maasai and Tanzanian Bantu populations (fr=<0.01), is absent from other Africans and Caucasians and only slightly more prevalent in Asian populations.

*CYP2D6*17* and *CYP2D6*29* are the most relevant functionally important alleles in African populations. *CYP2D6*17* frequencies ranged from 0.14 to 0.34, whereas *CYP2D6*29* was surprisingly lower in San, Maasai and Venda (fr<0.10). The study further confirmed that *CYP2D6*4*, which is Caucasian-specific, is generally low in African populations, with intermediate frequencies of 0.08, 0.08 and 0.09, in Maasai, Igbo and San, respectively.

The duplication allele *CYP2D6*2/2* was not determined here, due to difficulties in optimising the experimental conditions. Similar challenges were faced with *CYP2D6*5*, and only the San were genotyped for this allele, giving a frequency of 0.01. However, literature analysis indicated generally low frequencies of the duplication across all populations (Table 14), except Ethiopians (0.29).

This study marked the first time that *FMO3* variants were assessed in African populations. Three commonly known *FMO3* SNPs, affecting enzyme function, g.15167 G>A (E158K), g.18281 G>A (V257M) and g.21443 A>G (E308G), were analysed together with two other non-synonymous SNPs found in studies on African Americans, g.21599 T>C (L360P) and g.15089 G>C (D132H) (Lattard et al., 2003). The minor allele frequencies are shown in Table 14.

The K158 mutation was the most common variant allele in all populations. The San showed the lowest frequency of this allele (fr=0.33), whereas the frequency in the other African groups ranged from 0.42 to 0.52, similar to African Americans (fr=0.48) (Hao et al., 2007). The allele was found at similar frequencies in Caucasians (fr=0.43-0.44), while prevalent in Asians at only half that level (fr=0.19-0.23) (Cashman et al., 2001; Hao et al., 2007).

The prevalence of the M257 variant ranged from 0.01 to 0.05 in African populations. In an African American study, it was found at a frequency of 0.07, comparable with Caucasian data. In Asians however, this allele is twice as prevalent, with frequencies of 0.15 in Japanese and 0.20 in Chinese populations.

The G308 variant was not detected in Luo and Maasai, whilst its frequency ranged from 0.01 to 0.02 in the other African ethnic groups. This is significantly lower than the prevalence of this allele in Caucasians (fr=0.22-0.23) and Asians (fr=0.15-0.21).

The H132 variant is obviously African-specific and was not reported in Caucasians or Asians. It was found at significantly higher frequencies (0.05–0.08) amongst ethnic groups from west Africa (Hausa, Igbo, Yoruba) than in the groups from eastern (Kikuyu, Luo, Maasai) and southern (San, Shona, Venda) Africa (0.01–0.04) ($p<0.01$ in both cases), with no difference between the latter two regions. The P360 variant was not detected in any population.

Linkage disequilibrium was investigated for all five *FMO3* variants. The H132 mutation did not appear concurrently with any other variant in the inferred haplotypes, and the P360 variant was not found in any haplotype. It appears that the G308 mutation was detected in combination with the K158 variant at frequencies of 0.005-0.016 in Africans. This is much lower than in Caucasians and Asians, where frequencies as high as 0.22 were found in North Europeans and 0.2 in Japanese (Mao et al., 2009).

GSTM1 and *GSTT1* were genotyped for the presence or absence of the gene, and it was found that the frequency of both homozygous deletions is generally lower in Africans (fr=0.16-0.45) than in Caucasians or Asians, whose prevalence is mostly above 0.5. Only Caucasian frequencies of homozygous *GSTT1* deletions (fr=0.20-0.21) were in line with Africans and significantly lower than in Asian populations (fr=0.44-0.60).

The most common *NAT2* alleles in this study were *NAT2*5* (fr=0.20-0.42) and *NAT2*6* (fr=0.08-0.33), in line with Caucasian studies, but contrasting Asians, where only *NAT2*6* has a similar prevalence (fr=0.19-0.31), while *NAT2*5* is much rarer (fr=0.02-0.06).

The frequencies of *NAT2*7* and *NAT2*14* were obtained from the re-sequencing data and ranged from 0.01 to 0.06 and 0.03 to 0.14, respectively. These *NAT2*7* frequencies are in agreement with other African (fr=0.02-0.04) and Caucasian (fr=0.02-0.04) populations, whereas this allele appears more prevalent in Asian populations (fr=0.10-0.16). *NAT2*14* seems to be African-specific and was so far not found in Caucasians or Asians.

4.4. Population differentiation and relatedness

Population differences in SNP and allele frequencies were discovered as described above. The importance of these differences was analysed with the aim of determining their contribution to population differentiation and relatedness. Datasets from the above analyses were subjected to various statistic models as shown in Figure 5.

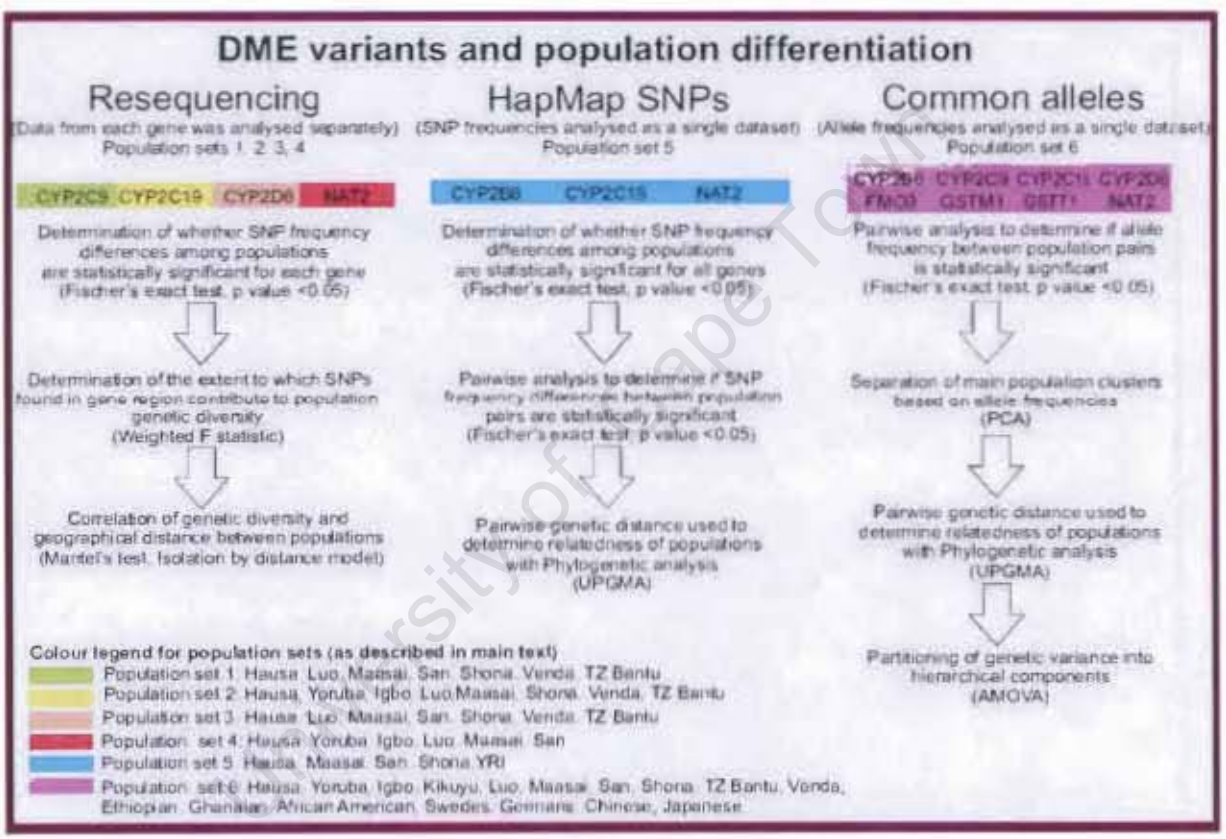


Figure 5: Data analysis for population differentiation and relatedness

4.4.1. Population variation in SNPs from re-sequencing analysis

Frequencies of SNPs detected by re-sequencing of CYPs and *NAT2* are shown in Tables 8-11. All of these SNPs were in HW equilibrium ($p>0.05$) (data not shown). Fischer's exact test indicated that there was significant inter-population variation for *CYP2C19* ($p=0.0097$),

CYP2D6 ($p<0.0001$) and *NAT2* ($p=0.00469$), whereas *CYP2C9* frequencies did not differ significantly ($p=0.94$). Weighted (average) Fst values revealed that the overall SNP variation in these gene loci plays a role in population differentiation for all genes except *CYP2D6*, displaying a negative value (-0.048; Table 10).

The fixation index Fst informs on the degree of differentiation but does not indicate the pattern of variation. Mantel's test for isolation by distance was applied to see if geographical distance between populations is correlated with SNP frequency diversity of the four re-sequenced genes. Converted pair-wise Fst distances (Fst/1-Fst) were plotted against pair-wise ln of geographical distances, and the summary statistics are shown in Table 15. *CYP2C9* showed some correlationbetween genetic differentiation and geographical distance ($r=0.0220$), which proved statistically significant ($p=0.011$). *CYP2C19* and *NAT2* showed modest correlations ($r=0.0040$ and $r=0.0089$, respectively), while negligible correlation was observed for *CYP2D6* ($r=0.00019$). As none of these correlations were statistically significant ($p>0.05$), only in case of *CYP2C9* geographical distance could explain genetic differentiation ($r^2=0.235$). *CYP2C19* and *NAT2* indicate much less determination of genetic differentiation by geographic distance, and the $r^2<0.001$ for *CYP2D6* suggests that this gene's pattern of variation cannot be explained by geography at all (Table 15).

Table 15: Genetic diversity in *CYP* and *NAT2* loci

A. Genetic differentiation (Fst)

CYP2C9	Maasai	Hausa	San	Shona	Venda	TZB
Hausa	0.0253					
San	0.0026	0.0221				
Shona	0.0128	0.065	0.0073			
Venda	0.0302	0.0922	0.0131	-0.025		
TZB	0.0247	0.0778	0.0176	0.0089	-0.0046	
Luo	-0.0099	-0.0059	-0.0215	-0.0043	0.0022	0.0009

CYP2C19	Shona	Venda	TZB	Hausa	Yoruba	Igbo	Luo
Venda	-0.0273						
TZB	-0.0168	-0.0308					
Hausa	0.0359	0.0181	0.0236				
Yoruba	-0.0026	-0.0149	0.0067	0.0019			
Igbo	-0.0151	-0.0079	-0.0126	0.0338	0.0149		
Luo	0.0436	0.0279	0.0128	0.0026	0.038	0.0462	
Maasai	0.0213	-0.0018	-0.0086	-0.0124	0.0038	0.0266	-0.0118

CYP2D6	Hausa	Yoruba	Igbo	Luo	Maasai	Shona
Yoruba	-0.031					
Igbo	0.0307	0.0028				
Luo	-0.0328	-0.0209	-0.0047			
Maasai	-0.0188	-0.0076	0.0791	-0.0064		
Shona	0.0144	0.0126	0.1171	0.0233	-0.0136	
Venda	-0.0472	-0.0173	0.0004	-0.0315	-0.0194	0.0278

NAT2	Hausa	Yoruba	Igbo	Luo	Maasai
Yoruba	-0.003				
Igbo	-0.0087	0.0126			
Luo	-0.0148	-0.0126	-0.0202		
Maasai	0.0113	0.0701	0.0009	0.0019	
San	0.0183	0.0118	0.0478	0.0149	0.0702

TZB=Tanzanian Bantu; Negative Fst indicates no role of the loci in the genetic differentiation of two populations; Pair-wise Fst values are shown for each gene locus.

B. Geographic distances (kilometres)

	Shona	Venda	TZB	Hausa	Yoruba	Igbo	Luo	Maasai
Venda	702							
TZB	1513	2171						
Hausa	4129	4595	3992					
Yoruba	4085	4449	4229	721				
Igbo	3613	4007	3736	740	495			
Luo	1874	2573	948	3067	3360	2877		
Maasai	2009	2710	795	3341	3673	3195	350	
San	449	426	1955	4173	4044	3595	1384	2550

TZB=Tanzanian Bantu; Distances obtained by estimating great circle distances (see Materials and Methods).

Table 15: Genetic diversity in *CYP* and *NAT2* gene loci *(continued)*

C. Isolation by distance statistics (Mantel's test)

Gene	r	p	r^2
CYP2C9	0.0220	0.011	0.235
CYP2C19	0.0040	0.315	0.011
CYP2D6	0.00019	0.78	<0.001
NAT2	0.0089	0.116	0.061

Isolation by distance is correlation of genetic distance (Fst/1-Fst) with ln of geographical distance; r=correlation coefficient, represents the slope of the regression line, illustrating the relationship between genetic diversity and geographic distance; p=probability value, proving the significance of the correlation; r^2=coefficient of determination, describing the proportion of genetic variation that can be explained by geographic distance.

4.4.2. Population variation in HapMap SNPs

In Table 13, frequencies of HapMap SNPs in the Yoruba (YRI) population were shown for comparison. All SNPs were in HW equilibrium ($p>0.05$), except for rs1801279 ($p<0.0097$), likely attributable to experimental error (data not shown).

Five of the *NAT2* variants (bold in Table 13: rs1801280, rs721399, *rs7832071*, *rs4646246*, rs1799929) showed significant allele frequency differences between the Hausa, Maasai, San and Shona (Table 13). Population pair-wise comparisons among the four populations revealed that Maasai allele frequencies were significantly different from the other populations (Table 16A). However, when *NAT2* SNPs were removed from the comparison, the difference in frequencies was no longer significant between population pairs (Table 16B).

Pair-wise genetic distances were plotted from SNP frequencies of the different populations, and a phylogenetic tree was constructed (Figure 6). Yoruba (YRI) appears distant from Hausa, San and Shona, while Maasai shows the largest separation from all the other populations.

Table 16: p-values for differentiation test of four ethnic groups based on HapMap SNP frequencies

A	Shona	San	Maasai
San	0.1855		
Maasai	**<0.0001**	**<0.0001**	
Hausa	0.79817	0.09936	**0.00055**

B	Shona	San	Maasai
San	0.54515		
Maasai	0.20120	0.14995	
Hausa	0.057503	0.29549	0.66486

Overall population differentiation was estimated using Fisher's exact test. **A**: using all SNP frequencies; **B** using SNP frequencies excluding NAT2; Bold=statistically significant ($p<0.05$).

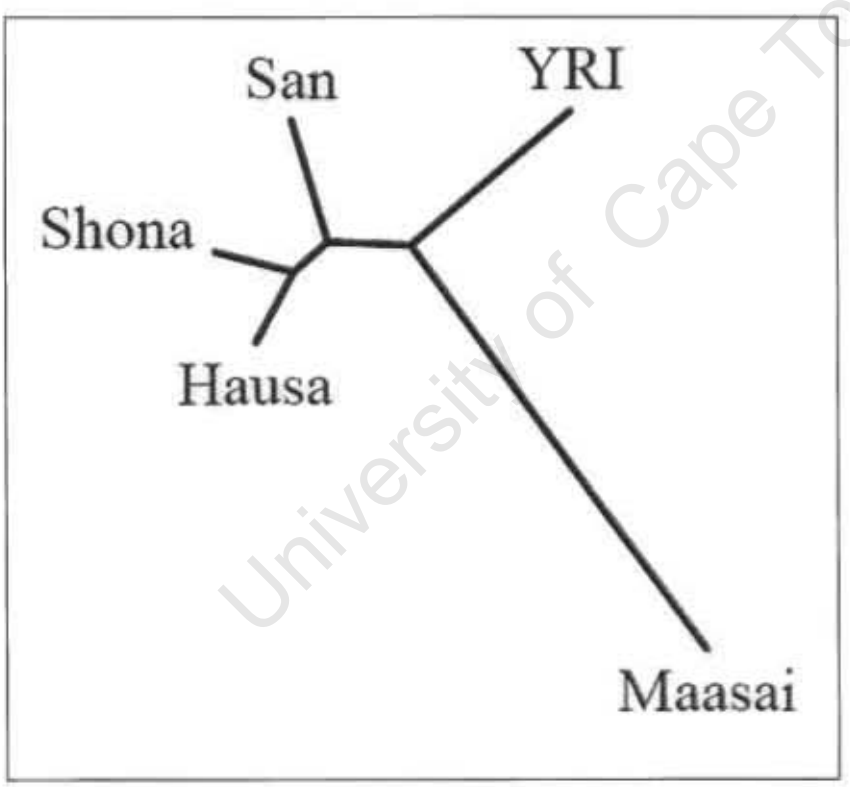


Figure 6: UPGMA tree calculated based on freqencies of 32 HapMap SNPs from *CYP2B6*, *CYP2C9*, *CYP2C19* and *NAT2* in four African populations, compared to HapMap Yoruba from Ibadan (YRI).

4.4.3. Population variation in common alleles

Population differentiation was assessed, based on frequencies of commonly known alleles (see Table 14). p-values of pair-wise comparisons of populations are given in Table 17, showing statistically significant differences ($p<0.05$) of all African populations versus Caucasians (Swedes, Germans) and Asians (Chinese, Japanese, Koreans) as well as of Caucasians and Asians versus each other. Based on the same frequency data in Table 14, the possibility of population clustering was explored. As this and downstream phylogenetic analysis require complete data sets, cases of lacking frequency data in literature were substituted with estimates based on average values of closest related populations. For example, *CYP2D6*3* was found absent in all African populations and therefore, regarded absent in African Americans. Average values for *NAT2* alleles were obtained from comparisons of various published African population data.

Principal Component Analysis (PCA) revealed distinct clustering of African, Caucasian and Asian populations (Figure 7A). The African cluster was mainly determined by the frequencies of *CYP2B6*6, CYP2D6*17, CYP2D6*29* and *NAT2*14,* which are significantly more prevalent in Africans than in the other two world populations (Figure 7B). The Caucasian cluster was largely defined by *CYP2D6*3, CYP2D6*4* and *CYP2D6*9*, which are almost exclusive to this population. In addition, *CYP2C9*2* and *CYP2C9*3* would be determinants of the Caucasian cluster (see Table 14), but these alleles were not included in the PCA analysis. The Asian cluster was mostly based on the frequencies of *CYP2C19*2*, *CYP2C19*3* and *CYP2D6*10*. Alleles located close to the zero cross section had little effect on clustering of populations.

Table 17: Pairwise comparison of populations (p-values based on allele frequencies from Table 14)

	Kikuyu	Luo	Maasai	Igbo	Yoruba	Hausa	San	Shona	Venda	TZB	Gha	Eth	AfA	Swe	Ger	Chi	Jap
Luo	0.994																
Maasai	**0.024**	0.088															
Igbo	**0.006**	0.186	**0.006**														
Yoruba	0.505	0.508	0.226	0.173													
Hausa	0.057	**0.038**	**0.011**	0.078	0.901												
San	**0.002**	**0.004**	**0.000**	**0.000**	**0.020**	**0.000**											
Shona	0.987	0.917	**0.014**	0.073	0.775	**0.022**	**0.000**										
Venda	0.840	0.878	0.193	0.363	0.705	0.251	**0.023**	0.938									
TZB	0.913	0.998	0.018	0.333	0.730	0.058	**0.001**	0.914	0.786								
Gha	0.724	0.922	**0.004**	0.054	0.694	0.017	0.142	0.886	0.681	0.964							
Eth	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.007**	**0.013**	**0.000**						
AfA	**0.040**	0.157	**0.002**	0.036	0.053	**0.015**	**0.015**	**0.013**	0.736	0.099	0.268	**0.000**					
Swe	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**				
Ger	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.999			
Chi	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**		
Jap	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.905	
Kor	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.687	0.803

TZB=Tanzanian Bantu; Gha=Ghanaian, Eth=Ethiopian; AfA=African American; Swe=Swede; Ger=German, Chi=Chinese, Jap=Japanese; Kor=Korean; Bold=statistically significant ($p<0.05$).

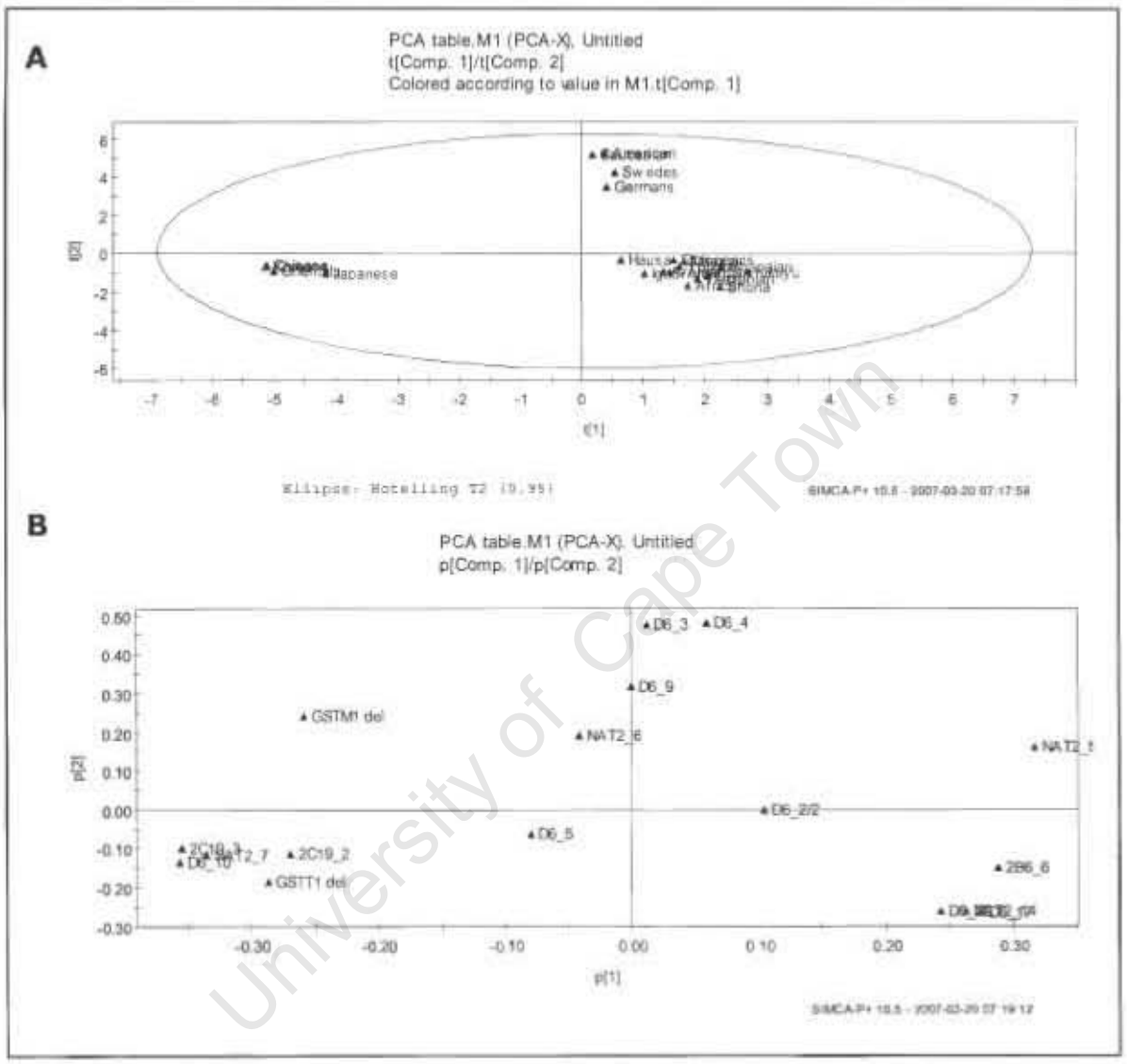


Figure 7: PCA analysis using allele frequencies of DME genes in Africans, Caucasians and Asians **A:** Separation of populations showing continental clusters **B:** Allelic determinants of cluster distribution.

UPGMA phylogenetic analysis was based on all populations in Table 14. As expected, three distinct clusters were observed, clearly separating Africans from Asians and Caucasians (Figure 8A). Further analysis of African populations showed clear separation of Ethiopians and San from a single sub-cluster formed by all other Africans (Figure 8B).

The hierarchy of variation **among and within** the three main world population clusters (Africans, Asians and Caucasians) was estimated using Analysis of Molecular Variance (AMOVA), based on allele frequencies in Table 14. Four levels were analysed: the world population clusters and various combinations of African ethnic groups represented for by A: countries, B: geographic regions, and C: ethno-linguistic classifications (Table 18).

AMOVA showed that variation mostly occurs within populations (>=97%). In contrast, variance was only 2.87% among world population clusters and negligible among ethnic groups within population clusters (0.11%).

Assessing African populations, among country variance was very low (0.03%), whereas the variation among ethnic groups within countries was slightly higher (0.22%). Inter-geographical regions had very low variance (0.01) while ethnic groups within geographical regions showed slightly higher variance (0.17%). Finally, ethno-linguistic families showed no variance amongst each other, but marginal variance was observed among ethnic groups within ethno-linguistic classes (0.9%).

Assessments of the fixation index Fst confirmed this hierarchy of variation, showing that only 2.971% of the total variation exists

among world populations, and even lower values were obtained for African populations when split by countries (0.249%), geographical regions (0.180%) or ethno-linguistic families (0.104%). All Fst values were shown to be statistically significant (P<<0.00001).

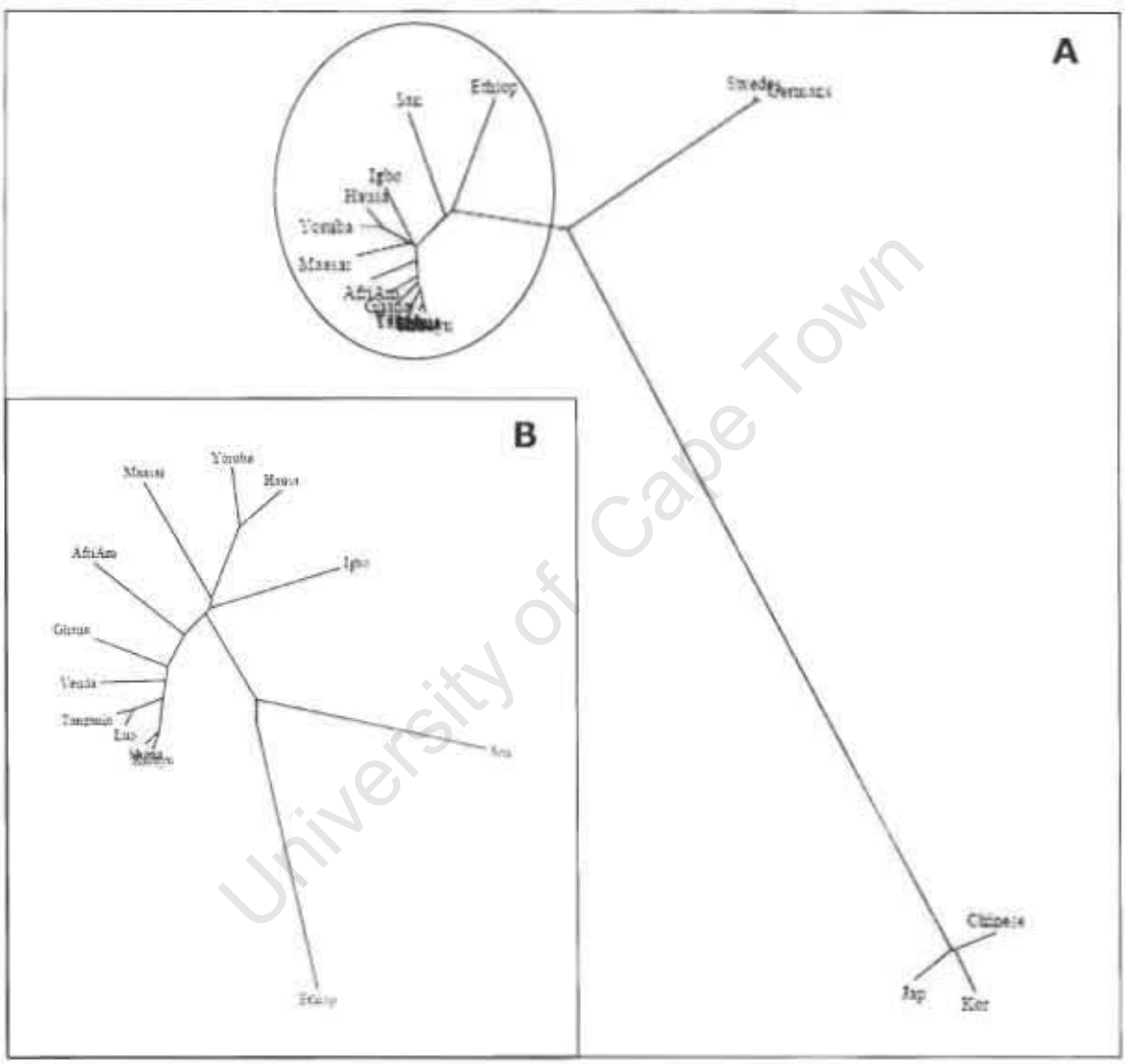


Figure 8: Un-rooted UPGMA tree showing population clusters based on frequencies of commonly known alleles **A:** All populations **B:** Africans; Tanzania=Tanzanian Bantu; Ethiop=Ethiopians, Afri Am=African Americans.

Table 18: AMOVA analysis on allele frequencies from Table 14

Source of variation	df	Sum of squares	Var	% variation
All population clusters*				
Among world population clusters	2	44.047	0.01328 Va	2.87
Among ethnic groups within population clusters	18	10.714	0.00049 Vb	0.11
Within world populations	6253	2812.158	0.44973 Vc	97.03
Total	6273	2866.92	0.4635	
Fst=0.02971 p<<0.00001				
African populations**				
A. Among countries	7	5.507	0.00012 Va	0.03
Among ethnic groups within countries	5	3.795	0.00101 Vb	0.22
Within country groups	3767	1700.351	0.45138 Vc	99.75
Total	3779	1709.653	0.45251	
Fst=0.00249 p<<0.00001				
B. Among geographical region groups***	2	1.448	0.00004 Va	0.01
Among ethnic groups within geographical regions	9	6.062	0.00078 Vb	0.17
Within geographical region groups	3459	1560.108	0.45103 Vc	99.82
Total	3470	1567.619	0.45184	
Fst=0.00180 p<<0.00001				
C. Among ethno-linguistic family groups**	2	1.126	0 Va	0
Among ethnic groups within ethno-linguistic families	8	4.603	0.00042 Vb	0.09
Within ethno-linguistic groups	3222	1455.045	0.4516 Vc	99.91
Total	3232	1460.775	0.45201	
Fst=0.00104 p<<0.00001				

*World population clusters: *Africans* (Kikuyu, Luo, Maasai, Hausa, Igbo, Yoruba, San, Shona, Venda, TZ Bantu, Ghanaian, Ethiopian, African American); *Asians* (Chinese, Japanese, Koreans); *Caucasians* (Swedes, Germans);
**excludes African Americans; A: African country groups as in Table 1, plus Ethiopians (Ethiopia), Ghanaians (Ghana); B: African geographical region as in Table 1, plus Ethiopians and Ghanaians in eastern and western Africa, respectively; C: African ethno-linguistic classes as in Table 1; df=degrees of freedom (n-1)(according to Arlequin v3.11); Var=Variance components.

4.5. Building pharmacogenetics resources in Africa

In order to achieve a widespread analysis of pharmacogenetic markers on the continent, samples were collected from geographically varied African ethnicities. The results were analysed and recorded towards establishing a data resource for pharmacogenetics studies in African populations (Figure 9).

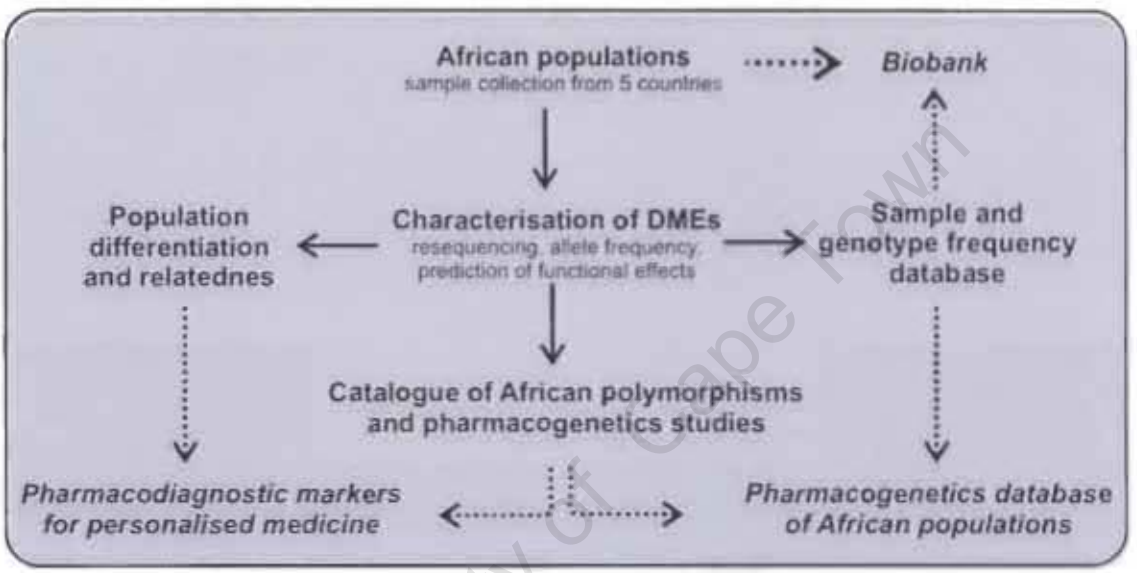


Figure 9: Resource building for pharmacogenetics in Africa.

4.5.1. Demographics of sample collection

Samples were collected from three main African regions, i.e. north-western (Nigeria), central-eastern (Kenya, Tanzania) and southern Africa (South Africa, Zimbabwe), thereby creating the most diverse collection of samples for pharmacogenetic studies in Africa so far. The population demographics of these ethnic groups was analysed in order to estimate the level of representation by the samples.

Hausa (Afro-Asiatic), Igbo and Yoruba (Niger Congo) represent 69% of the Nigerian population, hence data in this study would be representative of over 90 million Nigerians. Kikuyu (Niger Congo), Luo

(Nilotic), Maasai (Nilotic) and Mixed Tanzanian Bantu (Niger Congo) data would be representative of approximately 50 million people in Kenya and Tanzania of east Africa. In addition, Luo and Maasai ethnic groups are also found in Uganda, Sudan and Rwanda.

Both Shona and Venda belong to the Southern Bantu group. Samples collected from Zimbabwe were mainly of the Shona ethnicity, which is the main ethnic group representing 90% (10 million) of the Zimbabwe population.

The San population belongs to a small settlement in the Southern part of Zimbabwe bordering Botswana. DNA samples from the Venda of South Africa were from an old collection gathered from Limpopo province of the northern part of South Africa bordering Zimbabwe. Venda constitute a minority of the South African population (1 million people).

Therefore, based on population demographics, and for determining the general pharmacogenetic landscape, this collection and the data generated in the various analyses should be representative for at least 150 million people in Africa.

4.5.2. Sample and genotype database

A database was created to record genotype results of each sample analysed in this study. The database can be queried for sample location and genotypes defined by a particular test (Figures 10). Functions for basic statistical calculations such as minor allele frequencies or testing for HW consistence are included (Figure 11). Results can be uploaded to other programs for further analysis and presentation.

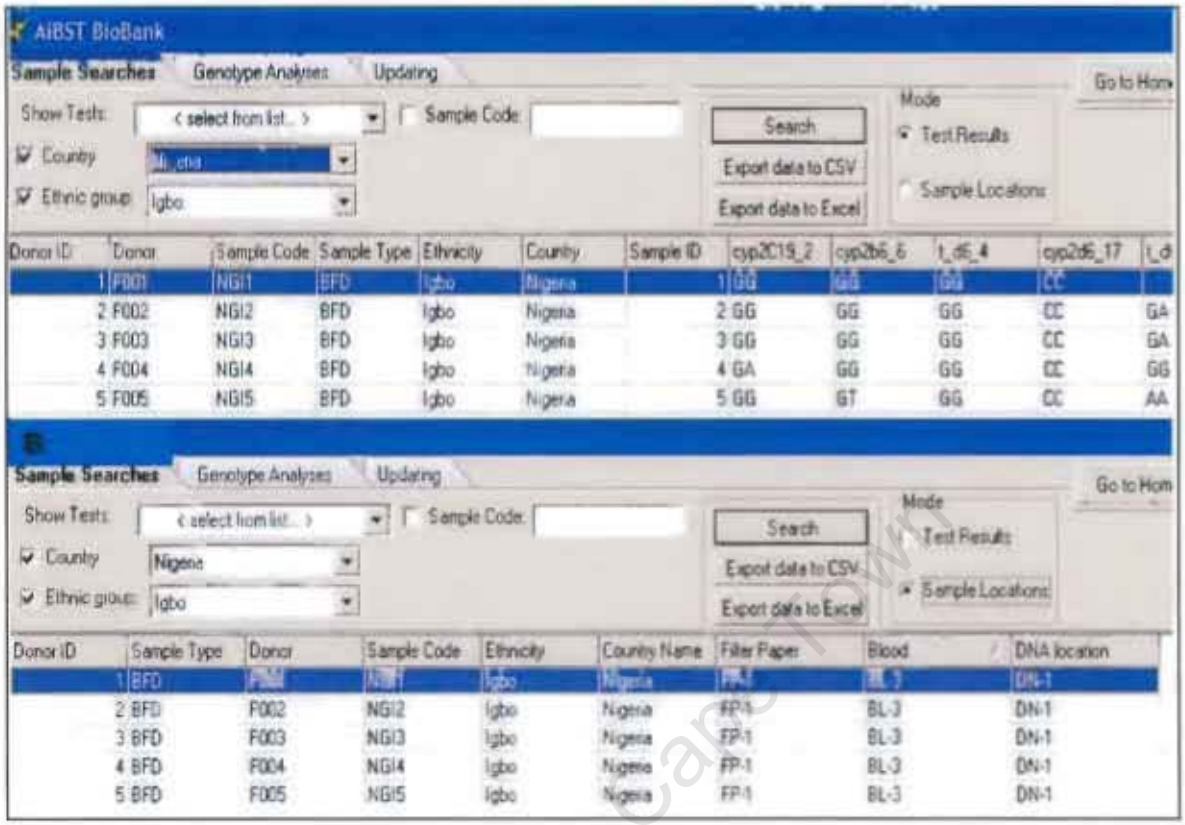


Figure 10: Sample and genotype database: Master table and general search options. **A:** Output of sample ID and genotype information, e.g. BFD means that the sample is available as blood, filter spots and DNA; **B:** Output of sample location information.

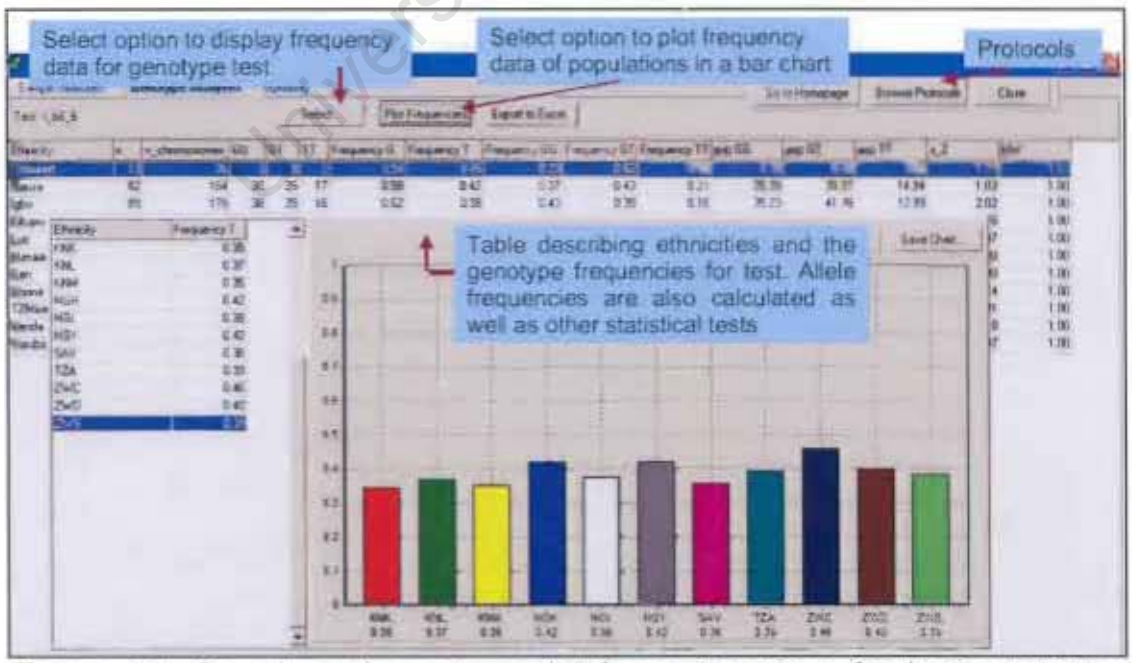


Figure 11: Sample and genotype database: Functions for basic statistical calculations.

4.5.3. Catalogue of African polymorphisms

The newly established catalogue of African polymorphisms features a total of 180 genetic variants (169 SNPs, 2 CNVs, 9 indels), analysed in this study, as shown in Table 19A and summarised in Table 20. Whereas the highest number of polymorphisms is in *CYP2D6*, both *GSTM1* and *GSTT1* were only analysed for their deletion allele. As the catalogue also contains polymorphisms found in African populations before, it gives a full account of all polymorphisms of the eight DME genes that were characterised in Africans to date.

A ranking system (Table 19B) was employed to categorise the levels of occurrence of these polymorphisms in African populations, deducing their importance for pharmacogenetic epidemiological studies. Whereas only three SNPs were confirmed to be rare, more than 50% of polymorphisms were found in the low or rare/low categories (Table 19C). Generally, most SNPs were located in introns. Non-synonymous SNPs were mostly in the low and low/rare categories.

4.5.4. Pharmacodiagnostic kit

The analysis was narrowed down to SNPs and CNVs that may affect protein expression and activity, hence be important for phenotypic outcome in individuals carrying them (Table 21). Where impact on enzyme expression and function is characterised, substrate drugs and the availability of phenotype data for Africans are indicated. Although some non-synonymous SNPs are shown to have no impact on enzyme function, these may be important in the context of functionally important haplotypes. In total, 60 polymorphisms from the eight DME genes are recommended to be included in a pharmacodiagnostic kit for the future determination of ADMET profiles in Africans.

Table 19: Catalogue of SNPs analysed in this study. **A.** Allele frequency ranking

Gene:reference accession ID	mRNA pos.	SNP	mRNA feature	effect	db SNP	Ha	Yo	Ig	Lu	Ma	Ki	Sa	Sh	Ve	TZB	Frequency ranking
2B6: NM_000767	8988	T>C	intron		rs3786547	0.44				0.36		0.36	0.39			high
2B6: NM_000767	**15638**	**G>T**	**exon 4**	**Q172H (*6)**	**rs3745274**	**0.42**	**0.42**	**0.38**	**0.37**	**0.35**	**0.34**	**0.40**	**0.38**	**0.36**	**0.39**	**high**
2B6: NM_000767	**18060**	**G>A**	**exon 5**	**K262R**	**rs2279343**	**0.40**				**0.43**		**0.47**	**0.42**			**high**
2B6: NM_000767	18979	C>T	intron		rs2306606	0.49				0.36		0.40	0.40			high
2B6: NM_000767	24329	C>T	intron		rs7259965	0.40				0.38		0.34	0.36			high
2C9: NM_000771	50056	A>T	intron		rs1934969	0.67			0.50	0.50		0.40	0.24	0.13	0.32	high
2C19:NM_000769	583	A>T	intron		rs10509676	0.46				0.51		0.50	0.50			high
2C19:NM_000769	12122	A>G	intron		rs7916649	0.50	0.37	0.29	0.58	0.50			0.25	0.28	0.33	high
2D6: M33388	214	G>C	intron		rs1080995	0.50	0.41	0.27	0.38	0.45			0.57	0.17	0.43	high
2D6: M33388	221	C>A	intron		rs1080996	0.50	0.41	0.27	0.38	0.45			0.64	0.30	0.43	high
2D6: M33388	223	C>G	intron		rs1080997	0.50	0.41	0.27	0.36	0.45			0.56	0.20	0.30	high
2D6: M33388	227	T>C	intron		rs1080998	0.50	0.41	0.27	0.36	0.45			0.50	0.25	0.50	high
2D6: M33388	232	G>C	intron		rs1080999	0.50	0.41	0.30	0.36	0.50			0.70	0.25	0.75	high
2D6: M33388	233	A>C	intron		rs1080999	0.50	0.41	0.27	0.38	0.45			0.67	0.25	0.50	high
2D6: M33388	245	A>G	intron		rs1081000	0.50	0.41	0.27	0.31	0.45			0.70	0.25	0.67	high
2D6: M33388	746	C>G	intron		nrs				0.36				0.40	0.31	0.50	high
2D6: M33388	843	T>G	intron		rs28371702	0.14	0.40	0.20	0.33	0.38			0.33	0.13	0.20	high
2D6: M33388	1661	G>C	exon 3	V136V	rs28371708	0.29	0.43	0.35	0.32	0.46			0.37	0.33	0.30	high
2D6: M33388	**2850**	**C>T**	**exon 6**	**R296C (*2)**	**nrs**	**0.54**	**0.50**	**0.50**	**0.55**				**0.63**	**0.44**	**0.65**	**high**
2D6: M33388	3384	A>C	intron		nrs	0.30	0.45	0.34	0.28	0.42			0.37	0.25	0.38	high
2D6: M33388	3584	G>A	intron		nrs	0.54	0.34	0.26	0.43	0.46			0.47	0.44	0.44	high
2D6: M33388	3790	C>T	splice site		nrs	0.53	0.34	0.26	0.44	0.54			0.47	0.44	0.44	high
2D6: M33388	**4180**	**G>C**	**exon 9**	**S486T**	**rs1135850**	**0.68**	**0.55**	**0.66**	**0.72**	**0.63**			**0.63**	**0.75**	**0.67**	**high**
2D6: M33388	4723	T>G	3'utr		nrs	0.63	0.57	0.73	0.57						0.25	high
FMO3: NM_006894	***c.472***	***G>A***	***exon 4***	***E158K***	***rs2266782***	***0.44***	***0.52***	***0.44***	***0.50***	***0.42***	***0.49***	***0.33***	***0.50***	***0.48***	***0.50***	***high***
GSTM1: NT_019273.18		***del***	***g***		***nrs***	***0.37***	***0.31***	***0.23***	***0.29***	***0.16***	***0.28***	***0.45***	***0.24***	***0.23***	***0.33***	***high***
GSTT1: NT_011520.11		***del***	***g***		***nrs***	***0.42***	***0.35***	***0.36***	***0.22***	***0.40***	***0.25***	***0.25***	***0.26***	***0.20***	***0.25***	***high***
NAT2: NM_000015	282	C>T	exon 2	Y94Y	rs1041983	0.40	0.44	0.55	0.44	0.38		0.29				high
NAT2: NM_000015	**341**	**T>C**	**exon 2**	**I114T (*5)**	**rs1801280**	**0.33**	**0.33**	**0.34**	**0.27**	**0.42**		**0.20**	**0.31**	**0.39**	**0.34**	**high**
NAT2: NM_000015	**590**	**G>A**	**exon 2**	**R197Q (*6)**	**rs1799930**	**0.32**	**0.33**	**0.29**	**0.30**	**0.25**		**0.20**	**0.24**	**0.22**	**0.21**	**high**
NAT2: NM_000015	**803**	**A>G**	**exon 2**	**K268R**	**rs1208**	**0.37**	**0.39**	**0.40**	**0.44**	**0.54**		**0.43**	**0.44**			**high**
NAT2: NM_000015		A>G	3'utr		rs721399	0.39				0.21		0.44	0.49			high
NAT2: NM_000015		C>T	3'utr		rs7832071	0.34				0.54		0.25	0.43			high

Table 19: Catalogue of SNPs analysed in this study. **A.** Allele frequency ranking *(continued)*

Gene:reference accesion ID	mRNA pos	SNP	mRNA feature	effect	db SNP	Ha	Yo	Ig	Lu	Ma	Ki	Sa	Sh	Ve	TZB	Frequency ranking
2B6: NM_000767	*3010*	*T>C*	*intron*		*rs8100458*	*0.22*				*0.25*		*0.24*	*0.09*			*medium*
2B6: NM_000767	*18634*	*G>A*	*intron*		*rs12721649*	*0.20*				*0.11*		*0.20*	*0.22*			*medium*
2C9: NM_000771	251	T>C	intron		rs9332104	0.08			0.10	0.27		0.12	0.22	0.17	0.11	medium
2C9: NM_000771	3411	T>C	intron		rs9332128	0.00			0.10	0.00		0.04	0.13	0.17	0.14	medium
2C9: NM_000771	9032	G>C	intron		nrs	0.00			0.10	0.14		0.15	0.15	0.13	0.14	medium
2C9: NM_000771	10311	A>G	intron		rs9332129	0.00			0.17	0.15		0.15	0.14	0.13	0.19	medium
2C9: NM_000771	**10535**	**A>G**	**exon 5**	**H251R (*9)**	**rs2256871**	**0.18**			**0.17**	**0.05**		**0.15**	**0.11**	**0.06**	**0.00**	**medium**
2C9: NM_000771	33349	A>G	intron		rs9332172	0.17			0.23	0.23		0.15	0.28	0.25	0.50	medium
2C9: NM_000771	33658	A>G	intron		rs9332174	0.13			0.14	0.27		0.12	0.20	0.19	0.10	medium
2C9: NM_000771	47593	T>C	intron		rs9332232	0.07			0.14	0.05		0.04	0.18	0.17	0.05	medium
2C19:NM_000769	-97	T>C	5'utr		rs4986894	0.13	0.18	0.33	0.07	0.08			0.20	0.11	0.15	medium
2C19:NM_000769	99	T>C	exon 1	P33P	rs17885098	0.05	0.08	0.15	0.18	0.15			0.17	0.17	0.30	medium
2C19:NM_000769	12662	A>G	splice site		rs12769205	0.16	0.20	0.33	0.09	0.12			0.27	0.22	0.20	medium
2C19:NM_000769	*13594*	*A>G*	*intron*		*rs4388808*	*0.23*				*0.22*		*0.09*	*0.11*			*medium*
2C19:NM_000769	*18427*	*G>A*	*intron*		*rs4304697*	*0.09*				*0.10*		*0.03*	*0.11*			*medium*
2C19:NM_000769	18911	A>G	intron		rs7088784	0.05	0.08	0.15	0.14	0.15			0.17	0.17	0.25	medium
2C19:NM_000769	**19154**	**G>A**	**exon 5**	**P227P (*2)**	**rs4244285**	**0.13**	**0.15**	**0.29**	**0.18**	**0.11**	**0.16**	**0.12**	**0.13**	**0.21**	**0.18**	**medium**
2C19:NM_000769	57740	G>C	intron		rs4417205	0.14	0.15	0.21	0.09	0.13			0.23	0.22	0.20	medium
2C19:NM_000769	80160	C>T	exon 7	V330V	rs3758580	0.13	0.18	0.25	0.07	0.08			0.17	0.17	0.15	medium
2C19:NM_000769	*81811*	*C>G*	*intron*		*rs11597626*	*0.20*				*0.22*		*0.33*	*0.24*			*medium*
2C19:NM_000769	87106	T>C	intron		rs4917623	0.38	0.13	0.21	0.31	0.23			0.03	0.06	0.15	medium
2C19:NM_000769	87522	C>T	intron		rs17885567	0.05	0.08	0.21	0.10	0.08			0.13	0.11	0.10	medium
2C19:NM_000769	89909	C>T	intron		rs12268020	0.22	0.31	0.17	0.09	0.17			0.20	0.17	0.20	medium
2D6: M33388	-175	G>A	5'utr		rs1080993	0.05	0.12	0.31	0.22	0.05			0.11	0.06	0.30	medium
2D6: M33388	-42	wt>insG	5'utr		rs28371695	0.19	0.35	0.13	0.21	0.05			0.20	0.13	0.10	medium
2D6: M33388	310	G>T	intron		rs28371699	0.00	0.27	0.07	0.25	0.31			0.25	0.00		medium
2D6: M33388	**1023**	**C>T**	**exon 2**	**T107I (*17)**	**rs28371706**	**0.20**	**0.20**	**0.22**	**0.22**	**0.17**	**0.33**	**0.22**	**0.34**	**0.24**	**0.18**	**medium**
2D6: M33388	1039	C>T	exon 2	F112F	rs1081003	0.00	0.13	0.13	0.04	0.00			0.00	0.13	0.05	medium
2D6: M33388	1067	T>G	intron		novel	0.13	0.13	0.19	0.04	0.08			0.07	0.19	0.10	medium
2D6: M33388	**1659**	**G>A**	**exon 3**	**V136M (*29)**	**rs1058164**	**0.11**	**0.10**	**0.28**	**0.24**	**0.08**	**0.14**	**0.02**	**0.17**	**0.06**	**0.25**	**medium**
2D6: M33388	**3183**	**G>A**	**exon 7**	**V338M (*29)**	**nrs**	**0.13**	**0.10**	**0.29**	**0.20**	**0.04**			**0.17**	**0.06**	**0.13**	**medium**
NAT2: NM_000015		*A>G*	*5'utr*		*rs4646246*	*0.28*				*0.09*		*0.15*	*0.32*			*medium*
NAT2: NM_000015	**191**	**G>A**	**exon 2**	**R64Q (*14)**	**rs1801279**	**0.03**	**0.08**	**0.11**	**0.14**	**0.09**		**0.09**	**0.14**	**0.11**	**0.13**	**medium**
NAT2: NM_000015	481	C>T	exon 2	L161L	rs1799929	0.25	0.14	0.34	0.27	0.46		0.14	0.26			medium
NAT2: NM_000015		*G>A*	*3'utr*		*rs4646247*	*0.27*				*0.27*		*0.14*	*0.24*			*medium*
NAT2: NM_000015		*T>C*	*3'utr*		*rs721398*	*0.27*				*0.27*		*0.19*	*0.19*			*medium*

Table 19: Catalogue of SNPs analysed in this study. **A.** Allele frequency ranking *(continued)*

Gene:reference accesion ID	mRNA pos	SNP	mRNA feature	effect	db SNP	Ha	Yo	Ig	Lu	Ma	Ki	Sa	Sh	Ve	TZB	Frequency ranking
2C9: NM_000771	3487	A>G	splice site		rs12769205	0.00			0.04	0.00		0.17	0.11	0.06	0.09	low/medium
2C9: NM_000771	50196	C>T	exon 9	A441A	rs2017319	0.04			0.14	0.05		0.04	0.20	0.17	0.05	low/medium
2C9: NM_000771	50434	C>T	3'utr		rs9332241	0.08			0.05	0.00		0.00	0.02	0.13	0.14	low/medium
2C19:NM_000769	90011	A>G	splice site		rs4451645	0.13	0.14	0.00	0.07	0.04			0.13	0.11	0.00	low/medium
2D6: M33388	3582	A>G	intron		nrs	0.08	0.11	0.11	0.09	0.04			0.00	0.13	0.00	low/medium
2D6: M33388	4401	C>T	3'utr		nrs	0.09	0.10	0.11	0.07	0.04			0.00	0.25	0.06	low/medium
2D6: M33388	4481	G>A	3'utr		nrs	0.12	0.08	0.03	0.11	0.21			0.23	0.07	0.19	low/medium
2D6: M33388	4656	wt>delA CA	3'utr		nrs	0.44	0.09	0.03	0.23	0.31			0.08	0.33	0.33	low/medium
2B6: NM_000767	*2613*	*G>A*	*intron*		*rs10426235*	*0.01*				*0.00*		*0.04*	*0.07*			*low*
2B6: NM_000767	*2854*	*G>A*	*intron*		*rs8192711*	*0.02*				*0.11*		*0.04*	*0.04*			*low*
2B6: NM_000767	*3097*	*T>C*	*intron*		*rs8192712*	*0.00*				*0.04*		*0.01*	*0.00*			*low*
2B6: NM_000767	***21018***	***C>T***	***exon 7***	***I328 T (*18)***	***rs28399499***											***low‡***
2C9: NM_000771	9069	G>A	intron		novel	0.00			0.05	0.05		0.00	0.00	0.00	0.05	low
2C9: NM_000771	9451	T>C	intron		rs17443251	0.00			0.05	0.00		0.00	0.00	0.00	0.06	low
2C19:NM_000769	188	G>A	intron		rs17881883	0.00	0.00	0.03	0.00	0.00			0.07	0.11	0.05	low
2C19:NM_000769	12306	G>A	intron		rs17878649	0.00	0.00	0.03	0.08	0.04			0.10	0.06	0.10	low
2C19:NM_000769	12637	C>T	intron		novel	0.03	0.10	0.00	0.00	0.04			0.00	0.06	0.00	low
2C19:NM_000769	18229	T>A	intron		rs17884938	0.06	0.03	0.05	0.07	0.00			0.10	0.00	0.00	low
2C19:NM_000769	57453	G>C	intron		novel	0.00	0.00	0.04	0.03	0.00			0.03	0.00	0.00	low
2C19:NM_000769	57512	A>G	intron		novel	0.03	0.05	0.08	0.03	0.04			0.00	0.00	0.05	low
2C19:NM_000769	57567	A>T	intron		novel	0.00	0.00	0.04	0.03	0.00			0.07	0.11	0.10	low
2C19:NM_000769	57575	T>C	intron		novel	0.00	0.00	0.04	0.03	0.00			0.03	0.00	0.00	low
2C19:NM_000769	57637	wt>delG	intron		novel	0.03	0.05	0.08	0.07	0.08			0.10	0.00	0.10	low
2C19:NM_000769	**80161**	**G>A**	**exon 7**	**V331I**	**rs3758581**	**0.03**	**0.03**	**0.00**	**0.00**	**0.04**			**0.00**	**0.00**	**0.00**	**low**
2C19:NM_000769	80629	T>A	intron 7		novel	0.03	0.05	0.00	0.00	0.05			0.00	0.00	0.00	low
2C19:NM_000769	**87290**	**T>C**	**exon 8**	**R410C (*13)**	**rs17879685**	**0.00**	**0.00**	**0.04**	**0.03**	**0.00**			**0.03**	**0.00**	**0.00**	**low**
2C19:NM_000769	87313	A>C	exon 8	G417G	rs17886522	0.03	0.05	0.08	0.07	0.04			0.10	0.00	0.00	low
2C19:NM_000769	87475	G>C	intron		rs17880188	0.08	0.13	0.00	0.07	0.04			0.11	0.06	0.00	low
2C19:NM_000769	89578	T>A	intron		rs12779363	0.00	0.00	0.06	0.04	0.00			0.03	0.00	0.00	low
2C19:NM_000769	90533	C>T	3'utr		novel	0.00	0.00	0.05	0.04	0.00			0.04	0.00	0.00	low
2D6: M33388	-150	C>T	5'utr		nrs	0.09	0.03	0.00	0.00	0.00			0.00	0.00	0.00	low
2D6: M33388	-85	T>C	5'utr		nrs	0.04	0.00	0.00	0.00	0.00			0.03	0.00	0.00	low
2D6: M33388	**77**	**G>A**	**exon 1**	**R26H (*43)**	**rs28371696**	**0.04**	**0.00**	**0.03**	**0.00**	**0.00**			**0.03**	**0.00**	**0.05**	**low**
2D6: M33388	**100**	**C>T**	**exon 1**	**P34S (*10)**	**rs1065852**	**0.13**	**0.07**	**0.10**	**0.06**	**0.05**			**0.06**	**0.12**	**0.04**	**low**
2D6: M33388	654	C>T	intron		novel		0.00	0.00	0.07				0.08	0.07	0.07	low

Table 19: Catalogue of SNPs analysed in this study. **A.** Allele frequency ranking *(continued)*

Gene:reference accesion ID	mRNA pos	SNP	mRNA feature	effect	db SNP	Ha	Yo	Ig	Lu	Ma	Ki	Sa	Sh	Ve	TZB	Frequency ranking
2D6: M33388	**1716**	**G>A**	**exon 3**	**E155K (*45)**	**rs28371710**	**0.08**	**0.00**	**0.00**	**0.09**	**0.04**			**0.00**	**0.17**	**0.00**	**low**
2D6: M33388	**1846**	**G>A**	splice site	**splicing defect (*4)**	**nrs**	**0.03**	**0.08**	**0.08**	**0.04**	**0.04**	**0.01**	**0.09**	**0.02**	**0.03**	**0.05**	**low**
2D6: M33388	**1863_ 1864**	**ins(TTT CGC CCC)X2**	**exon 4**	**174_175 ins(FRP) X2**	**nrs**	**0.00**	**0.00**	**0.00**	**0.04**	**0.08**			**0.00**	**0.00**	**0.00**	**low**
2D6: M33388	2575	C>A	exon 5	P267P	nrs				0.05				0.03	0.22	0.00	low
2D6: M33388	2602	G>T	exon 5	L276L	novel				0.05				0.00	0.06	0.00	low
2D6: M33388	2661	G>A	intron		nrs				0.05				0.03	0.11	0.05	low
2D6: M33388	2760	T>A	intron		novel				0.00				0.10	0.06	0.00	low
2D6: M33388	3254	T>C	exon 7	H361H	rs2743457	0.09	0.00	0.00	0.07	0.08			0.00	0.13	0.00	low
2D6: M33388	3561	G>C	intron		novel	0.00	0.00	0.00	0.02	0.00			0.00	0.06	0.06	low
2D6: M33388	3707	G>A	intron		nrs	0.00	0.00	0.03	0.00	0.00			0.03	0.00	0.00	low
2D6: M33388	**3853**	**G>A**	**exon 8**	**E410K (*27)**	**nrs**	**0.00**	**0.00**	**0.00**	**0.06**	**0.04**			**0.00**	**0.00**	**0.06**	**low**
2D6: M33388	4033	C>T	splice site		novel	0.00	0.00	0.00	0.02	0.00			0.00	0.06	0.06	low
2D6: M33388	4394	wt>delAG	3'utr		novel	0.00	0.00	0.00	0.00	0.00			0.10	0.06	0.00	low
2D6: M33388		**deletion**		***5**												**low‡**
2D6: M33388		**multip**		**2D6XN**												**low‡**
FMO3: NM_006894	***394***	***G>C***	***exon 4***	***D132H***	***rs12072582***	***0.05***	***0.07***	***0.08***	***0.02***	***0.03***	***0.01***	***0.02***	***0.04***	***0.03***	***0.03***	***low***
FMO3: NM_006894	***769***	***G>A***	***exon 6***	***V257M***	***rs1736557***	***0.04***	***0.01***	***0.01***	***0.05***	***0.04***	***0.05***	***0.01***	***0.02***	***0.02***	***0.02***	***low***
FMO3: NM_006894	***923A***	***A>G***	***exon 7***	***E308G***	***rs2266780***	***0.01***	***0.01***	***0.01***	***0.00***	***0.00***	***0.01***	***0.01***	***0.02***	***0.02***	***0.01***	***low***
NAT2: NM_000015	**809**	**T>C**	**exon 2**	**I270T**	**novel**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**		**0.13**				**low**
NAT2: NM_000015	**838**	**G>A**	**exon 2**	**V280M**	**nrs**	**0.06**	**0.03**	**0.05**	**0.03**	**0.00**		**0.00**				**low**
NAT2: NM_000015	**857**	**G>A**	**exon 2**	**G286E (*7)**	**rs1799931**	**0.03**	**0.03**	**0.03**	**0.03**	**0.04**		**0.01**	**0.06**	**0.05**	**0.03**	**low**
2C9: NM_000771	-375	T>C	5'utr		rs9332103	0.04				0.00		0.00	0.00			rare/low
2C9: NM_000771	3499	T>A	splice site		rs9332121	0.00			0.00	0.00		0.00	0.03	0.00	0.00	rare/low
2C9: NM_000771	***3627***	***G>A***	***exon 4***	***R150H (*8)***												***rare/low‡***
2C9: NM_000771	8963	T>C	intron		nrs	0.00			0.05	0.00		0.00	0.00	0.00	0.00	rare/low
2C9: NM_000771	10347	T>C	intron		novel	0.00			0.00	0.00		0.00	0.03	0.00	0.00	rare/low
2C9: NM_000771	**10601**	**wt>delA**	**exon 5**	**K273 fs (*6)**	**nrs**	**0.04**			**0.00**	**0.00**		**0.00**	**0.00**	**0.07**	**0.00**	**rare/low**
2C9: NM_000771	42415	C>T	intron		novel	0.00			0.00	0.06		0.00	0.00	0.00	0.00	rare/low

Table 19: Catalogue of SNPs analysed in this study. **A.** Allele frequency ranking *(continued)*

Gene:reference accesion ID	mRNA pos	SNP	mRNA feature	effect	db SNP	Ha	Yo	Ig	Lu	Ma	Ki	Sa	Sh	Ve	TZB	Frequency ranking
2C9: NM_000771	42469	T>C	intron		rs9332197	0.00			0.00	0.05		0.00	0.00	0.00	0.00	rare/low
2C9: NM_000771	**42519**	**T>C**	**exon 7**	**I327T (*31)**	**rs57505750**	**0.04**			**0.00**	**0.00**		**0.00**	**0.00**	**0.00**	**0.00**	**rare/low**
2C9: NM_000771	***42542***	***C>T***	***exon 7***	***R335W (*11)***												***rare/low?***
2C9: NM_000771	**42619**	**G>C**	**exon 7**	**D360E (*5)**	**rs28371686**	**0.00**			**0.00**	**0.00**		**0.00**	**0.02**	**0.06**	**0.00**	**rare/low**
2C9: NM_000771	47545	A>T	intron		rs9332230	0.00			0.00	0.00		0.00	0.00	0.00	0.05	rare/low
2C9: NM_000771	47639	C>T	intron		rs2298037	0.00			0.00	0.00		0.00	0.03	0.06	0.00	rare/low
2C9: NM_000771	50125	C>T	intron		novel	0.00			0.00	0.00		0.00	0.00	0.00	0.05	rare/low
2C9: NM_000771	**50294**	**A>G**	**exon 9**	**N474S**	**novel**	**0.00**			**0.05**	**0.00**		**0.00**	**0.00**	**0.00**	**0.00**	**rare/low**
2C9: NM_000771	50298	A>T	exon 9	G475G	rs1057911	0.00			0.00	0.00		0.00	0.00	0.00	0.05	rare/low
2C9: NM_000771	**50341**	**G>T**	**exon 9**	**V490F (*32)**	**novel**	**0.00**			**0.00**	**0.00**		**0.00**	**0.00**	**0.00**	**0.05**	**rare/low**
2C9: NM_000771	50413	C>T	3'utr		rs9332240	0.00			0.05	0.00		0.00	0.03	0.00	0.00	rare/low
2C9: NM_000771	50501	C>T	3'utr		rs9332243	0.00			0.05	0.00		0.00	0.03	0.00	0.00	rare/low
2C19:NM_000769	**55**	**A>C**	**exon 1**	**I19L (*15)**	**rs17882687**	**0.00**	**0.00**	**0.00**	**0.05**	**0.00**			**0.00**	**0.00**	**0.05**	**rare/low**
2C19:NM_000769	183	T>C	intron		rs17882201	0.00	0.00	0.00	0.00	0.00			0.00	0.06	0.00	rare/low
2C19:NM_000769	231	A>C	intron		novel	0.00	0.00	0.00	0.01	0.00			0.00	0.00	0.00	rare/low
2C19:NM_000769	**12460**	**G>C**	**exon 2**	**E92D**	**rs17878459**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**			**0.03**	**0.00**	**0.00**	**rare/low**
2C19:NM_000769	12607	wt>insC	intron		novel	0.00	0.00	0.00	0.03	0.04			0.00	0.00	0.00	rare/low
2C19:NM_000769	**12690**	**G>A**	**exon 3**	**V113I**	**novel**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**			**0.00**	**0.06**	**0.00**	**rare/low**
2C19:NM_000769	**12784**	**G>A**	**exon 3**	**R144H (*9)**	**rs17884712**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**			**0.00**	**0.06**	**0.00**	**rare/low**
2C19:NM_000769	**17869**	**G>T**	**exon 4**	**R186P (*22)**	**novel**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**			**0.00**	**0.00**	**0.06**	**rare/low**
2C19:NM_000769	**17948**	**G>A**	**exon 4**	**W212X (*3)**	**rs4986893**	**0.01**	**0.00**	**0.00**	**0.00**	**0.04**	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**rare/low**
2C19:NM_000769	18207	G>A	intron		novel	0.00	0.00	0.00	0.02	0.00			0.00	0.00	0.06	rare/low
2C19:NM_000769	18254	T>C	intron		novel	0.03	0.03	0.00	0.00	0.00			0.00	0.00	0.00	rare/low
2C19:NM_000769	18818	T>C	intron-premiRNA (has-mir-139)		novel	0.00	0.03	0.00	0.00	0.00			0.00	0.00	0.00	rare/low
2C19:NM_000769	19076	T>C	splice site		novel	0.00	0.00	0.00	0.00	0.00			0.03	0.00	0.00	rare/low
2C19:NM_000769	19332	G>A	intron-premiRNA (has-mir-448)		novel	0.00	0.00	0.00	0.00	0.08			0.00	0.00	0.00	rare/low
2C19:NM_000769	57678	T>G	intron		rs28399511	0.00	0.00	0.00	0.00	0.04			0.00	0.00	0.00	rare/low
2C19:NM_000769	57989	G>C	intron		novel	0.00	0.00	0.00	0.02	0.00			0.10	0.00	0.00	rare/low

Table 19: Catalogue of SNPs analysed in this study. **A.** Allele frequency ranking *(continued)*

Gene:reference accesion ID	mRNA pos	SNP	mRNA feature	Effect	db SNP	Ha	Yo	Ig	Lu	Ma	Ki	Sa	Sh	Ve	TZB	Frequency ranking
2C19:NM_000769	87422	A>G	intron 8		novel	0.03	0.00	0.00	0.00	0.08			0.00	0.00	0.00	rare/low
2C19:NM_000769	**90209**	**A>C**	**exon 9**	**X491C; 26 extra aa (*12)**	**nrs**	**0.00**	**0.00**	**0.00**	**0.04**	**0.00**			**0.03**	**0.00**	**0.00**	**rare/low**
2C19:NM_000769	90301	C>T	3'utr		novel	0.01	0.00	0.00	0.00	0.00			0.00	0.00	0.00	rare/low
2C19:NM_000769	90302	C>T	3'utr		novel	0.00	0.00	0.00	0.04	0.00			0.03	0.00	0.00	rare/low
2D6: M33388	**82**	**C>T**	**exon 1**	**R28C (*21)**	**nrs**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**			**0.00**	**0.00**	**0.05**	**rare/low**
2D6: M33388	1006	C>T	exon 2	R101R	novel	0.00	0.00	0.00	0.00	0.00			0.00	0.00	0.05	rare/low
2D6: M33388	**1608**	**G>A**	**exon 3**	**V119M (*70)**	**novel**	**0.00**	**0.00**	**0.00**	**0.02**	**0.00**			**0.00**	**0.00**	**0.00**	**rare/low**
2D6: M33388	1621	G>T	exon 3	R123L	novel	0.00	0.00	0.00	0.00	0.00			0.00	0.00	0.05	rare/low
2D6: M33388	1866	C>T	exon 4	N175N	nrs	0.00	0.00	0.00	0.02	0.00			0.00	0.00	0.00	rare/low
2D6: M33388	1869	T>C	exon 4	G176G	nrs	0.00	0.03	0.00	0.00	0.00			0.00	0.00	0.00	rare/low
2D6: M33388	1998	T>C	exon 4	F219F	novel			0.00	0.00				0.03	0.00	0.05	rare/low
2D6: M33388	2988	G>A	intron		nrs				0.00				0.03	0.00	0.00	rare/low
2D6: M33388	**3259_3260**	**wt>insTG**	**exon 7**	**375 fs (*42)**	**nrs**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**			**0.03**	**0.00**	**0.00**	**rare/low**
2D6: M33388	3397	C>A	intron		novel	0.00	0.00	0.00	0.00	0.00			0.00	0.06	0.00	rare/low
2D6: M33388	3721	wt>delGT	intron		nrs	0.00	0.03	0.00	0.00	0.00			0.00	0.00	0.00	rare/low
2D6: M33388	**4057**	**G>A**	**exon 9**	**G445E**	**novel**	**0.00**	**0.00**	**0.00**	**0.04**	**0.00**			**0.00**	**0.00**	**0.00**	**rare/low**
NAT2: NM_000015	**403**	**C>G**	**exon 2**	**L135V**	**nrs**	**0.00**	**0.03**	**0.00**	**0.03**	**0.00**		**0.00**				**rare/low**
NAT2: NM_000015	**472**	**A>C**	**exon 2**	**I158L**	**novel**	**0.00**	**0.00**	**0.00**	**0.03**	**0.00**		**0.00**				**rare/low**
NAT2: NM_000015	**589**	**C>T**	**exon 2**	**R197X**	**novel**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**		**0.01**				**rare/low**
NAT2: NM_000015	**641**	**C>T**	**exon 2**	**T214I**	**novel**	**0.00**	**0.00**	**0.00**	**0.03**	**0.00**		**0.00**				**rare/low**
NAT2: NM_000015	**683**	**C>T**	**exon 2**	**P228L**	**nrs**	**0.00**	**0.00**	**0.00**	**0.03**	**0.00**		**0.00**				**rare/low**
NAT2: NM_000015	**766**	**A>G**	**exon 2**	**K256E**	**nrs**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**		**0.03**				**rare/low**
2C9: NM_000771	**3608**	**C>T**	**exon 3**	**R144C (*2)**	**rs1799853**	**0.00**			**0.00**	**0.00**		**0.00**	**0.00**	**0.00**	**0.00**	**rare**
2C9: NM_000771	**42614**	**A>C**	**exon 7**	**I359L (*3)**	**rs1057910**	**0.00**			**0.00**	**0.00**		**0.00**	**0.00**	**0.00**	**0.00**	**rare**
FMO3: NM_006894	***1079***	***T>C***	***exon 8***	***L360P***	***rs28363581***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***rare***

mRNA pos=mRNA position; SNP frequencies were obtained from re-sequencing analysis, except for bold italics, which were obtained from RFLP, Taqman or Sequenom genotyping; †not analysed in this study and ranking based on allele frequencies from literature sources; all bold: SNPs proposed for pharmacodiagnostic kit for African populations; novel=found in this study; nrs=no reference number assigned; del=deletion; g=gene; multip=multiple copies of gene; Ha:Hausa; Yo:Yoruba; Ig:Igbo; Lu: Luo; Ma: Maasai; Ki: Kikuyu; Sa: San; Sh:Shona; Ve:Venda; TZB: TZ Bantu; ns=non-synonymous, s=synonymous; UTR=untranslated region; CNV=copy number variation.

B. Ranking key for Table 19A

<0.01	rare
0.01-0.10	low
0.10-0.30	medium
>0.30	high

C. Summary of frequency ranking for Table 19A

feature	high	medium	low/medium	low	low/rare	rare
exon ns	8	5		14	25	3
exon s	2	5	1	4	5	
splice site	1	1	2	2	2	
Intron	17	19	1	22	20	
UTR	3	6	4	4	5	
CNV	2			2		
Total	***33***	***36***	***8***	***48***	***57***	***3***

Table 20: Summary of catalogue (from Table 19A)

Number of ethnic groups/populations	10
Number of samples	997
Number of genes	8
Polymorphisms analysed	180
CYP2B6	10
CYP2C9	33
CYP2C19	52
CYP2D6	58
FMO3	5
GSTM1	1
GSTT1	1
NAT2	20

Table 21: Polymorphisms proposed for pharmacodiagnostic kit based on genes in this study

Gene:ref seq. ID	Effect	Enzyme	Drugs studied in Africans
CYP2B6: NM_000767	Q172H (*6)	decreased expression/ activity	efavirenz, nevirapine
	K262R (part of *6 or *18 haplotype)	decreased expression/ activity	efavirenz, nevirapine
	I328T (*18)	decreased expression/ activity	efavirenz, nevirapine
CYP2C19:NM_000769	P227P (*2) (introduction of cryptic splice site)	no activity	S-mephenytoin, omeprazole
	V331I	ne	
	R410C (*13)	ns	
	I19L (*15)	ns	
	E92D (part of *2 haplotype)	no activity	na
	V113I	ns	
	R144H (*9)	decreased activity	na
	R186P (*22)	ns	
	W212X (*3)	no protein	na
	X491C; 26 extra aa (*12)	unstable protein	na
CYP2C9: NM_000771	H251R (*9)	ns	
	R150H (*8)	decreased activity	losartan
	K273 fs (*6)	no protein	losartan
	I327T (*31)	ns	
	R335W (*11)	decreased activity	losartan
	D360E (*5)	decreased	losartan
	N474S	ns	
	V490F (*32)	ns	
	R144C (*2)	decreased activity	losartan/warfarin
	I359L (*3)	decreased activity	losartan/warfarin
CYP2D6: M33388	R296C (*2 haplotype)		
	S486T (*2 haplotype)		
	T107I (*17)	decreased activity	dextromethorphan/debrisoquine
	V136M (*29 haplotype)	decreased activity	dextromethorphan/debrisoquine
	V338M (*29 haplotype)	decreased activity	dextromethorphan/debrisoquine
	R26H (*43)	ns	

Table 21: Polymorphisms proposed for pharmacodiagnostic kit for genes in this study *(continued)*

Gene:ref seq. ID	Effect	Enzyme	Drugs studied in Africans
CYP2D6:M33388	P34S (*10)	decreased activity	dextromethorphan/debrisoquine
	E155K (*45)	ns	
	splicing defect (*4)	no protein	dextromethorphan/debrisoquine
	174_175ins(FRP)X2 (*40)	no activity	dextromethorphan/debrisoquine
	E410K (*27)	ne	
	deletion (*5)	no expression	dextromethorphan/debrisoquine
	(multiplication) 2D6XN	increased activity	dextromethorphan/debrisoquine
	R28C (*22)	ns	
	V119M (*70)	ns	
	375 fs (*42)	no protein	na
	G445E	ns	
FMO3: NM_006894	E158K	decreased activity	na
	D132H	decreased activity	na
	V257M	decreased activity	na
	E308G	decreased activity	na
	L360P	decreased activity	na
GSTM1: NT_019273.18	null	no expression	association with cancer
GSTT1: NT_011520.11	null	no expression	association with cancer
NAT2: NM_000015	I114T (*5)	decreased activity	isoniazid
	R197Q (*6)	decreased activity	isoniazid
	K268R	ne	
	R64Q (*14)	decreased activity	isoniazid
	I270T	ns	
	V280M	ns	
	G286E (*7)	decreased activity	isoniazid
	L135V	decreased activity	na
	I158L	ns	
	R197X	ns	
	T214I	ns	na
	P228L	decreased activity	na

ns=not studied; ne=no effect on enzyme expression or activity; na=not studied in African population.

5. DISCUSSION

Drug metabolising enzymes (DME) cytochrome P450 CYP2B6, CYP2C9, CYP2C19 and CYP2D6, as well as FMO3, GSTM1, GSTT1 and NAT2 are involved in the metabolism of pharmaceutical drugs and have been implicated in cancers due to their role in detoxification of carcinogenic agents and oxygen radicals. Polymorphisms in their genes account for variation of plasma drug concentrations in patients and their diagnostic use is rapidly finding clinical applications globally (see Table 3). Characterising such genetic variants and mapping out their distribution in representative African populations of various ethno-linguistic classes (see Table 1) was the main aim of this study.

The re-sequencing and genotyping analysis of DME genes in Africans done here is the most extensive effort of its kind to date. Novel variants' possible effects on enzyme expression and function were predicted by bioinformatics, contributing to understand phenotypes in drug metabolism. Applying statistical models, the variants' prevalence was used to explore diversity and evolutionary relatedness of African populations. In order to help translating these results into personalised medicine in Africa, bioresource building and the advancement towards pharmacodiagnostic applications are discussed.

5.1. Discovery of novel SNPs in DME genes

SNPs account for much of the genetic variation in drug metabolism. So far, over 30 such variants have been reported for Cytochrome P450 *CYP2C9*, approximately 20 for *CYP2C19*, over 60 for *CYP2D6* (http://www.cypalleles.ki.se/), (Sim and Ingelman-Sundberg, 2006) and 19 for N-acetyltransferase *NAT2* (Hein et al., 2008). Whereas

Caucasian and Asian populations have been studied extensively, very little data exists on Africans. Here, novel SNPs were discovered by re-sequencing of DME genes in various African ethnicities.

The experimental strategy was based on the specifics of the CYP and NAT gene structures as well as cost efficiency considerations. Whereas *CYP2D6* is contained within a 5kb region, *CYP2C9* and *CYP2C19* span over 50kb and 91kb, respectively, and all three genes have nine exons each. Priority was placed on analysing exonic sequences due to the higher likelihood of linking amino acid changes with functional importance. The *NAT2* gene is about 13kb in size, of which the 100bp exon 1 is non-coding. Therefore, only the coding exon 2 of 900bp was analysed. Splice site junctions spanning at least 50bp from each side of the exons as well as UTR sequences were included in the analysis.

Gene-specific primers enable the isolation of specific gene regions or whole genes by PCR amplification. This is the method of choice in the context of gene families such as the CYP genes that are highly homologous: more than 90% for *CYP2C9* and *CYP2C19*, and 95% for *CYP2D6* and the pseudogenes *CYP2D7* and *CYP2D8* (Kimura et al., 1989).

Isolating whole genes by long-range PCR is highly desirable and has been used to separate *CYP2D6* before (Gaedigk et al, 1999). However, due to difficulties in optimising this method, here the gene was fully covered by amplifying shorter *CYP2D6* fragments (up to 1.2kb), using primers and experimental conditions described previously (Masimirembwa et al., 1996). In addition, a modified nucleotide (deaza-GTP, see Materials and Methods) was used, owing to the high

GC content of *CYP2D6*, causing strong base–base interactions, which lead to superstructures and consequently to regions with higher melting temperatures lowering amplification efficiency by DNA polymerases (Jung et al., 2002).

Long-range PCR also proved problematic when tested for genotyping of common *CYP2D6* alleles (see 3.4.1). In most cases, the method was inconsistent and non-reproducible, amplification products were scarce, yielding faint bands on agarose gels. Only for genotyping *CYP2D6*5* in the San population, long-range PCR could be used. This indicates that the quality of DNA samples, together with the optimisation of experimental conditions, such as calibration of PCR machines and use of high-fidelity, thermostable DNA polymerases (e.g. Pfu DNA polymerase), (Cline et al., 1996; Lundberg et al., 1991), seems to be a highly critical parameter for long-range PCR.

The Sanger sequencing method (Sanger et al., 1977) was optimised to read sequences of up to 600 bp in one direction. Due to economic limitations on re-sequencing of whole genes, the strategy used here did not allow for identification of copy number variants or SNPs in introns and other upstream and downstream regulatory regions. Variation screening by re-sequencing, covering whole genes at higher throughput and lower cost, will become available with new-generation sequencing technology in the near future (see 1.3.2.1).

5.1.1. Functional characterisation of mutations

Non-synonymous SNPs in coding regions (exons), causing an amino acid change in the enzyme protein, are of primary interest in pharmacogenomic studies. However, only a few amino acid changes may directly affect the function of the enzyme. Others may be

involved in expression mechanisms or be in linkage with other causal mutations. Predictions of functional effects are based on various levels of protein structure. First, amino acid chemistry, such as a change from glycine to tryptophan, may have an impact due to the different sizes and charges of these amino acids. Second, the substitution of/by proline may affect the secondary and tertiary structure of the protein. Third, predictions are based on conservation in the alignment of known sequences from the same protein families. For example, CYP substrate recognition regions are based on Gotoh's alignment (Gotoh, 1992). Recently, several human CYP proteins have been crystallised, including CYP2C9 and CYP2D6 (Rowland et al., 2006b; Williams et al., 2003a), and the crystal structure of NAT2 is now also available (Wu et al., 2007). In cases where crystal structures are not available (e.g. CYP2C19), homology modelling is used. It is assumed that such approximation is sufficiently accurate to predict functional effects on substrate recognition, binding and catalysis of reactions (de Graaf et al., 2007; Wang et al., 2007).

Polyphen (http://genetics.bwh.harvard.edu/pph; Ramensky et al., 2002) and SIFT (Ng and Henikoff, 2003) are currently the most popular prediction algorithms, deriving structural data from protein databases such as Protein Data Bank (PDB; www.rcsb.org) and SWALL (http://srs.ebi.ac.uk). Here, Polyphen was used for predicting functional effects of novel non-synonymous SNPs represented as PSIC scores (Table 12).

Amino acid changes with a PSIC score of less than 1, such as N474S in CYP2C9, V113I in CYP2C19, V119M in CYP2D6, I158L and I270T in NAT2, are assumed not to be involved in any functional sites and

predicted not to affect enzyme function. Although V119M in CYP2D6 is located close to the substrate recognition site, the change from valine to methionine may have little impact in terms of hydrophobicity. On the other hand, a PSIC score can be <1 and yet have an impact on enzyme function. For example, the known amino acid change R144H in CYP2C19*9 has a PSIC score of 0.503, yet is predicted to damage the enzyme. The replacement of arginine by histidine might change the hydrophobicity at a buried site and decrease enzyme activity in vitro (Blaisdell et al., 2002). Functional impact is of course position dependent. For example, it would be predicted that R410C in CYP2C19*13 has no impact on the enzyme (PSIC score=0.336), although the change is dramatic, from large size, basic (R) to medium size, uncharged (C). However, structural analysis suggests that the residue is on the surface and hence less likely to be involved in substrate recognition. This does of course not rule out the possibility of surface amino acid changes affecting the maintenance of the protein's globular structure.

Some changes with PSIC scores slightly above 1 may still have modest effects on enzyme function. For example, R123L in CYP2D6 (PSIC score=1.236), when aligned with Gotoh's sequences (Gotoh, 1992), was shown to be involved in the substrate recognition site SRS1 (Bapiro et al., 2002). The T214I change in NAT2 (PSIC score=1.257) seems to interfere with enzyme function because this residue is important for interaction with the co-enzyme A ligand, according to structural prediction (Wu et al., 2007). The amino acid change D360E in CYP2C9*5 appears to have a modest impact on enzyme function (PSIC=1.475) because both residues have similar physico-chemical properties, being medium size and acidic. But since this position is at

the core of the protein, the slight difference may result in a significant structural effect, hence the observed decrease in enzyme activity as found in vitro and in individuals carrying this SNP (Allabi et al., 2004; Dickmann et al., 2001).

The effect of the R186P change in CYP2C19 leads to a change in electrostatic charge and possibly geometry; hence, it is predicted to affect the protein dramatically, giving a high PSIC score (3.159). An equally high score (3.063) is obtained for G445E in CYP2D6 due to the change from small, uncharged glycine to acidic glutamic acid. This may impact this residue's possible interaction with close position 443, which is important for heme-ligand binding and therefore has a high probability of affecting enzyme function.

In conclusion, these predictions must be interpreted with caution and do not always correlate with in vitro/in vivo studies or clinical phenotype. Based on validated phenotypes, a full analysis of all CYP genes, using Polyphen and SIFT for predicting functional effects of non-synonymous SNPs, showed 70% prediction accuracy for these algorithms, yet there were some discrepancies (Wang et al., 2009). For example, T107I and V136M;V338M in CYP2D6*17 and CYP2D6*29, respectively, were predicted to have no impact on enzyme function by Polyphen and SIFT, yet have been shown to result in reduced enzyme activity in individuals carrying these alleles (Masimirembwa et al., 1996; Wennerholm et al., 2001). To overcome these limitations, more recent developments include theoretical measures to identify the residues that are critical for maintaining structural stability by assessing the consequences on the interaction network of single point mutations (Cheng et al., 2008).

5.1.2. SNPs affect mRNA processing

Although most past analyses focused on non-synonymous, coding SNPs, non-coding SNPs and coding synonymous SNPs can play important roles in gene expression. Cis-acting elements, such as promoter as well as 5′ and 3′ UTRs, exon/intron junctions and splice recognition sites contain important regulatory and other functional information. Intronic and UTR SNPs may be involved in disrupting sequences recognised by trans-acting factors that affect splicing, transcription or other mRNA processing events.

Therefore, exon/intron junctions and some UTRs were included in the analysis here. Information-based theory (see 3.6.4) was used to analyse novel synonymous SNPs, as well as intronic SNPs within the splice sites (−25 to +2 for exon acceptor sites and −3 to +6 for exon donor sites) of *CYP2C9*, *CYP2C19* and *CYP2D6*, but no significant effects on splice site recognition were found.

Such effects were studied in CYPs before. Whereas some defective splice site variants are well understood, for example *CYP2D6*4* (1846 G>A), which occurs at the zero acceptor position of exon 4 (Hanioka et al., 1990), functional indications are less clear if mutations lie further away from splice site junctions. Information theory analysis has been used to show how other intronic and synonymous mutations may contribute to splice site effects in CYP genes (Rogan et al., 2003). For example, the defective allele *CYP2C19*2* (19154 G>A) results in a synonymous mutation (P227P), yet it has been associated with poor metabolism of CYP2C19 drugs (de Morais et al., 1994; Ibeanu et al., 1998). Further investigations showed that this mutation introduces a

cryptic splice site forty nucleotides downstream, resulting in a truncated non-functional protein.

Recent molecular advances have revealed that microRNA recognition sequences (pre-miRNA) are involved in regulation of protein expression and may be important in pharmacogenomic variation (Passetti et al., 2009). Mutations in these sequences, as well as insertions of new pre-miRNA sequences, could affect enzyme expression. However, *CYP1B1* is the only CYP that has been found to be miRNA-regulated so far (Tsuchiya et al., 2006). In the present study, none of the SNPs introduced pre-miRNA sequences in the 3' UTRs, yet in *CYP2C19*, 18818 T>C in intron 4 and 19332 G>A in intron 5 introduced miRNA binding sites for has-mir-139 and has-mir-448, respectively (Table 9). However, since miRNA binding sites mostly act within 3' UTRs (Rajewsky, 2006), these mutations would not be expected to have functional effects.

5.1.3. Validation of functional effects of SNPs

DME gene polymorphisms, that cause changes in amino acid sequence, gene regulation or mRNA processing, affect amounts and/or structures of proteins, thereby shaping metabolic phenotype variability (Sadee and Dai, 2005). To validate functional effects of these polymorphisms, production and stability of mRNA and protein are measured, and catalytic activity of enzymes is assessed, using probe substrates. This data is then correlated with clinical studies.

Site directed mutagenesis and expression in *E.coli* has been widely used for CYP enzymes such as CYP2C9, CYP2C19, CYP2D6 (Blaisdell et al., 2002; Deeni et al., 2001; DeLozier et al., 2005; Yun et al., 2006). The CYP protein is often targeted to the plasma membrane using an N-

terminal membrane anchor (Yun et al., 2006a). However, as heme incorporation or protein folding can be problematic in *E.coli*, Baculovirus-mediated expression in insect cells has been used for functional characterisation of CYP SNPs (Zhang et al., 2009). NAT2 variants have been expressed in mammalian cells (Zang et al., 2007). Expressed recombinant proteins are characterised by *in vitro* enzyme kinetics in combination with high performance liquid chromatography (HPLC). For example, S-mephenytoin hydroxylation by CYP2C19 (Blaisdell et al., 2002) or dextromethorphan O-demethylation by CYP2D6 (Yu et al., 2001; Zhang et al., 2009) serve as enzyme function markers, allowing comparison of catalytic activities of recombinant mutant proteins .

In addition to limitations of *in silico* predictions and *in vitro* assays, it has been found that some amino acid changes in drug metabolising enzymes affect different substrates in different ways. Such substrate-dependent impact of polymorphisms has been observed for CYP2D6 (Bogni et al., 2005; Zhang et al., 2009). Hence, in addition to overall protein function, substrate affinity and binding properties should be assessed for functional relevance of polymorphisms. Drug substrates, such as anti-depressants metabolised by CYP2D6 and warfarin metabolised by CYP2C9, have been used to determine metabolic ratios and/or plasma concentrations for evaluating pharmacokinetic profiles in individuals (Kirchheiner and Rodriguez-Antona, 2009; Klein et al., 2009).

In summary, novel SNPs predicted *in silico* to have a high impact on protein structure and function are the main candidates for *in vitro* characterisation studies. Here, the following novel, non-synonymous

SNPs are proposed for further characterisation *CYP2C9*31* (I327T), *CYP2C9*32* (V490F); *CYP2C19*22* (R186P) and *CYP2D6* 4057 G>A (G445E).

5.2. Complex haplotype networks

5.2.1. Allele nomenclature and new alleles

Assigning SNPs to haplotypes allows to determine whether an SNP is likely to have an indirect impact on gene expression and enzyme function based on other SNPs in the same haplotype. Allele nomenclature of CYPs and NATs is built on determining the major detrimental SNP in a haplotype. In other words, only SNPs predicted with high probability to affect gene expression and/or enzyme function are assigned an allele name. Novel SNPs found in introns or representing synonymous changes are not assigned new allele names.

For example, *CYP2D6*17* is defined as 1023 C>T; 1661 G>C; 2850 C>T; 4180 G>C, causing amino acid changes T107I; R296C; S486T, whereas *CYP2D6*64* is described by -1426 C>T; -1235 A>G; -1000 G>A; 100 C>T; 310 G>T; 843 T>G; 1023 C>T; 1661 G>C; 2097 A>G; 2850 C>T; 3384 A>C; 3582 A>G; 4180 G>C; 4401 C>T; 4722 T>G, causing amino acid changes P34S; T107I; S486T. It appears that the SNPs 1023 C>T; 1661 G>C; 2850 C>T; 4180 G>C are present in both alleles, and this would have been determined by the haplotype prediction programme Haploview (Barrett et al., 2005b). The defining SNPs would be 1023 C>T for *CYP2D6*17* and both 100 C>T and 1023 C>T for *CYP2D6*64*. This shows that SNPs in one haplotype may be present in the context of other SNPs in another haplotype, creating complex haplotype networks.

Haplotypes were estimated for the novel non-synonymous SNPs in linkage with other known mutations. Whereas it was possible to assign allele names to *CYP2C9*31*, *CYP2C9*32*, *CYP2C19*22* and *CYP2D6*70* (Table 12), the prevalence of *CYP2C9* 50294 A>G (N474S), *CYP2C19* 12690 G>A (V113I), *CYP2D6* 1621 G>T (R123L) and *CYP2D6* 4057 G>A (G445E) was too low to confirm the haplotypes predicted for them. In addition, none of the novel *NAT2* SNPs seemed to be in linkage with other known polymorphisms.

5.2.2. HapMap SNPs

The complex haplotype network was further highlighted in an attempt to determine the extent to which haplotype blocks based on HapMap Yoruba of Ibadan (YRI) data can be transferred to other African populations. Due to the extensive linkage disequilibrium (LD) between neighbouring loci in the human genome, it is believed that a subset of the SNPs in a region (tagSNPs) can be selected to capture most of the remaining SNP variants (Service et al., 2007). It was shown that HapMap tagSNPs selected with $r^2 >= 0.8$ can capture more than 85% of the SNPs in populations from the same continental group (Xing et al., 2008). This would estimate the rate at which certain SNPs are in LD when analysed in different populations.

SNPs of *CYP2B6*, *CYP2C19* and *NAT2* presented in the haplotypes of HapMap YRI were genotyped in four populations (Hausa, Maasai, San and Shona), followed by haplotype reconstruction. However, due to limitations in the design of the experiment, it proved impossible to capture all SNPs defining the major haplotype blocks in these genes. Problems were caused by primer specificity in multiplex PCR reactions and the presence of extended regions of nucleotide repeats. In

addition, differences were found in SNP density and nature of haplotypes spanning them, based on diverse recombination and evolutionary events. Whereas the *CYP2C19* region is captured in a single block covering the whole coding region, this is not the case for *CYP2B6* and *NAT2*, for which much of the SNPs are carried in various haplotype blocks (Figure 12). This made it difficult to create the critical combinations of SNPs for analysis, hence the selection of SNPs does not reflect the true and complete picture of haplotype tagSNPs as would have been preferred.

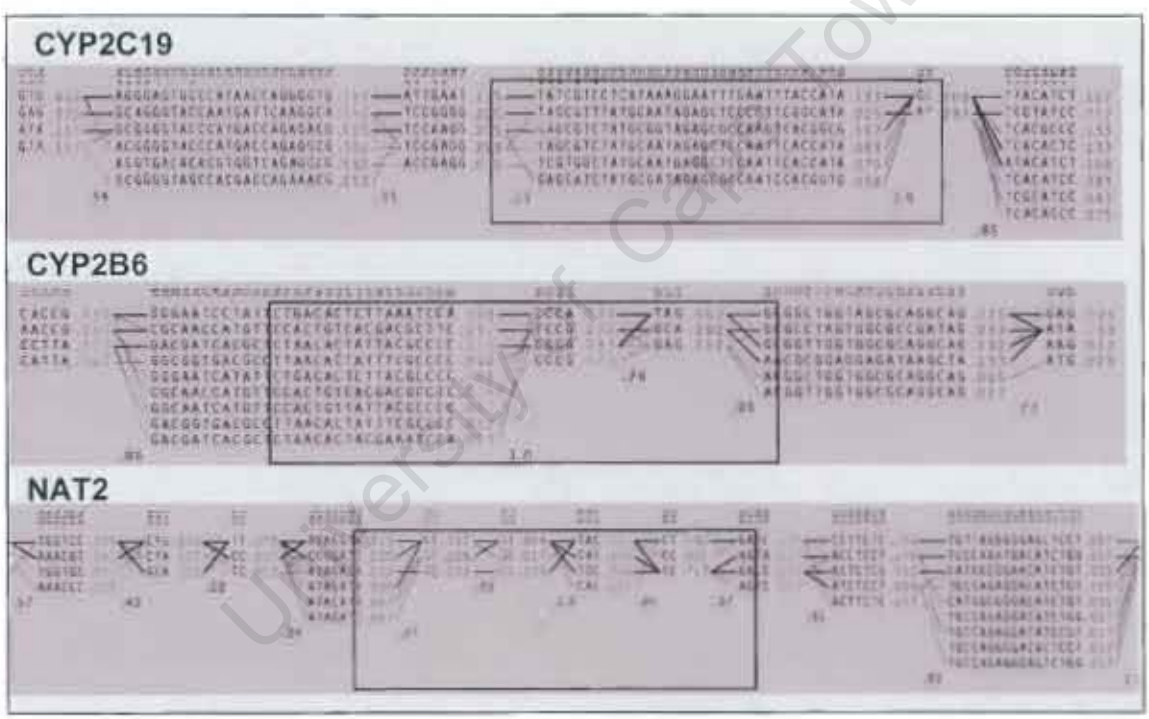


Figure 12: Complex pattern of haplotypes in CYP2C19, CYP2B6 and NAT2 gene regions; boxes indicate the haplotype blocks containing each gene. Figures were generated from Haploview using SNPs downloaded from HapMap (www.hapmap.org, frequency>5% in YRI). Gene regions in boxes: CYP2C19: Chromosome 10, positions 96,512,453 to 96,602,660; CYP2B6: Chromosome 19, positions 46,189,044 to 46,216,141; NAT2: Chromosome 8, positions 18,293,035 to 18,303,003.

These difficulties should be easier to overcome in the future, with more African populations, such as Maasai and Luhya of Webuye, and Maasai of Kinyawa, Kenya, having recently been added to the HapMap project (Manolio et al., 2008). As the African genome is complex, featuring lower levels of LD than other world populations (Campbell and Tishkoff, 2008), this is an important step towards better understanding of African haplotype characteristics.

5.3. Prevalence and clinical impact of DME polymorphisms

The low numbers of novel non-synonymous SNPs discovered here suggest that the most prevalent polymorphisms of functional importance had already been found in African populations. However, functionally important alleles can be of low prevalence and hence more difficult to find. Some view these as less important in realising their effects at a population level and therefore, less relevant in large clinical trials, where a general baseline prevalence is needed. However, on an individual level of diagnosis, rare mutations are most relevant and should definitely be considered for genotyping applications towards personalised medicine, including all mutations that have been validated for functional effects and annotated for phenotypical outcome. As the prevalence of variants varies between populations, certain SNPs or haplotypes that have been reported as prevalent and functionally important in other populations are rare or have not yet been detected in Africans. Here, genetic polymorphisms of *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *NAT2*, *GSTM1*, *GSTT1* and *FMO3* are discussed with respect to their potential clinical applications for drug

dosage and environmental exposure most relevant in African populations.

5.3.1. CYP2B6

*CYP2B6*6* (516 G>T) is the most commonly studied allele that has an impact on the pharmacokinetics of CYP2B6 substrate drugs. It is more prevalent in Africans (35-50%) than in Asians and Caucasians (15-25%). Functional studies showed that this allele displays remarkably lower catalytic activity and a significant decrease in protein expression (Hofmann et al., 2008).

Studies of efavirenz exposure levels in HIV/TB co-infected patients indicate that patients homozygous for the 516 G>T variant had higher plasma concentrations, over the minimum safe level of 4µg/ml. Pharmacokinetic modelling indicated that such patients might require as low as 300mg/day efavirenz, half the standard dose of 600mg/day. The data from the Zimbabwean patients also showed that irrespective of genotype, women had consistently higher efavirenz plasma levels than men (Nyakutira et al., 2008).

Another example is *CYP2B6*5*, which is detected in 14-25% of Caucasians and 8% of Africans and causes a significant decrease in protein expression and catalytic activity, but only in females (Lamba et al., 2003). Another SNP which has been associated with elevated plasma concentrations of efavirenz and nevirapine in Africans is 983 T>C (Wang et al., 2006), which can be found either alone as *CYP2B6*16* or in combination (785 G>A, 983 T>C) as *CYP2B6*18*. This SNP has a frequency of 4-7% in Africans and is absent in Caucasians and Asians (Wyen et al., 2008). In conclusion, this data emphasises the importance of ethnicity and gender in drug dosing.

5.3.2. CYP2C9

As one of the first official endorsement of pharmacogenomics for personalised medicine, the FDA has recommended genotyping of *CYP2C9*2* and *CYP2C9*3* in combination with *VKORC1* polymorphisms, intending a more targeted use of the anticoagulant warfarin (Thompson, 2007) and the prevention of excessive bleeding episodes (Higashi et al., 2002). In general, the effects of *CYP2C9* alleles have been studied extensively in Caucasians, including population-based pharmacokinetics–pharmacodynamics modelling, highlighting the importance of pharmacogenetics in drug use (Dickinson et al., 2007; Lindh et al., 2009; Stehle et al., 2008).

CYP2C9 variants *CYP2C9*2* (R144C) and *CYP2C9*3* (I359L) are the most common and occur at frequencies of 0.11 and 0.08, respectively, in Caucasians (Yasar et al., 1999). However, these alleles were not found in the African populations of this study, confirming their rare occurrence and low impact formerly reported in African Americans (frequency~0.01) (Kealey et al., 2007).

Whereas the *CYP2C9* genotype has been used to estimate some dose adjustments in the use of warfarin and phenytoin in Caucausians and Asians (Ohno et al., 2009; Stehle et al., 2008), the apparent absence or low frequency of these variants in African populations means that their plasma concentration variability of warfarin should be due to other genetic and/or environmental factors. This highlights the need to tailor pharmacodiagnostic tools for specific populations, based on baseline frequency of the variants of interest.

The most common non-synonymous *CYP2C9* allele detected by re-sequencing in this study was *CYP2C9*9* (H251R) (frequency=0.11). This variant is predicted to be damaging to enzyme function, yet phenotypic studies in African individuals have shown no effect on the metabolism of the anti-epileptic drug phenytoin. However, genotype-phenotype discrepancies of drug effects are common in African populations, probably due to unknown mutations or other genomic variations (Gaedijk et al., 2005). The overall distribution of slow metabolisers of phenytoin in Africans remains unclear due to limited data (Allabi et al., 2004). The prevalence of several low frequency alleles such as *CYP2C9*5* (D360E) and *CYP2C9*6* (K273fs) may explain such variation in African populations. *CYP2C9*5* causes impaired enzyme activity and *CYP2C9*6*, first found in African Americans, has been associated with phenytoin toxicity (Kidd et al., 2001). Other low frequency alleles such as *CYP2C9*8* (R150H) and *CYP2C9*11* (R335W) were also found to result in decrease in phenytoin metabolism (Allabi, 2005).

5.3.3. CYP2C19

*CYP2C19*2* (splicing defect) and *CYP2C19*3* (W212X) have been recommended as biomarkers for the administration of certain CYP2C19 substrates, including proton pump inhibitors and antidepressants (Furuta et al., 2007). The CYP2C19 slow metaboliser phenotype is detected in 2-4% of Caucasians and in about 20% of Asians, and these two variants account for 99% of slow metaboliser phenotypes in Asians (Goldstein et al., 1997; Ibeanu et al., 1998). In general, the clinical relevance of *CYP2C19* polymorphism seems to be more important for Asian populations compared to Caucasians and Africans.

*CYP2C19*2* is the most prevalent known defective *CYP2C19* variant in Africans, featuring an average frequency range of 0.1 to 0.3, highest in Igbo (Table 14), but considerably lower than in Asians (fr=0.51). *CYP2C19*3* is rare in African populations, consistent with earlier studies (Xie et al., 1999b), whereas it occurs at fr=0.08-0.12 in Asians. Here, this allele was observed at low frequency in Maasai and Hausa samples (Table 14), and one individual was identified as heterozygous in the Tanzanian population, in agreement with previous reports (Herrlin et al., 1998).

An earlier report shows that *CYP2C19*2* accounts for over 70% of slow metabolisers of S-mephenytoin in Africans (Masimirembwa et al., 1995). The missing 30% might be made up by *CYP2C19*3* and other low frequency variants such as *CYP2C19*12*, *CYP2C19*13* and *CYP2C19*15*, which would make these SNPs important contenders to include in genotyping panels for diagnostic purposes in African populations. Recently, *CYP2C19*17* has been discovered at a frequency of 0.18 in Ethiopians and Swedes, associated with increased enzyme activity (Allabi et al., 2004; Rudberg et al., 2008; Sim et al., 2006). Individuals carrying this variant may require a higher dosage in order to achieve the therapeutic effect of the drug omeprazole, due to estimations of 35% to 40% lower omeprazole plasma concentrations in *CYP2C19*17* homozygotes (Sim et al., 2006).

5.3.4. CYP2D6

Drugs affected by *CYP2D6* polymorphisms include debrisoquine, nortriptyline, metoprolol, fluoxetine and amitriptyline. In case of pro-drugs, metabolism leads to the pharmacologically active substance, as for codeine, which is O-demethylated to morphine by CYP2D6. For breast cancer treatment with tamoxifen, CYP2D6 status has been

reported to be an independent outcome predictor (Goetz et al., 2007; Tan et al., 2008).

The analysis of diverse African populations confirmed that the African-specific alleles *CYP2D6*17* (T107I) and *CYP2D6*29* (V136M, V338M) remain the most important functional SNPs for Africans' metabolism of CYP2D6 substrate drugs. Together with other less prevalent haplotypes, they explain why African populations generally have a larger number of intermediate metabolisers (~40%) compared with Caucasian populations (~15%) (Gaedigk et al., 2008).

Other defective *CYP2D6* alleles occuring in Africans include *CYP2D6*4* and *CYP2D6*10*. *CYP2D6*4* is a null allele, which is most common in Caucasians (frequency~0.2), responsible for over 70-90% of slow metabolisers, while *CYP2D6*10* produces an enzyme with reduced activity and is most common in Asians (frequency=0.43-0.51).

Clinically, the high frequency of intermediate metabolisers due to high prevalence of CYP2D6*17 and *29 could require that some narrow therapeutic index CYP2D6 substrate drugs need to be dosed at lower levels compared to Caucasians. Asians, who also have a higher prevalence of intermediate metabolisers due to the high frequency of CYP2D6*10, may administer lower doses of CYP2D6 substrate drugs compared to Caucasians (Kitada, 2003).

Other SNPs with significant impact on genotype-phenotype correlation, found in Africans, are *CYP2D6*36*, *CYP2D6*40*, *CYP2D6*45* and *CYP2D6*56*. The importance of these low-prevalence alleles is still difficult to ascertain due to isolated case studies.

Copy number variants (CNVs) of *CYP2D6*, which result in ultra-rapid or slow metabolisers, are well known. Multiplication of *CYP2D6* has been reported in southern European populations (Bernal et al., 1999; Scordo et al., 2004), and most significantly in Ethiopians, reaching frequencies as high as 0.29 (Aklillu et al., 1996). This is in contrast with generally lower frequencies (fr=0.02) in the rest of Africans. Whole gene deletions, causing slow metaboliser phenotypes, have been detected at more uniform frequencies across all populations (fr=0.03-0.06). In general, *CYP2D6* CNVs have significant consequences on toxicity and effectiveness of pharmaceutical treatment. Individuals heterozygous for a deletion polymorphism, such as *CYP2D6*5*, or a reduced-function allele, such as *CYP2D6*10* or *CYP2D6**17, may become slow metabolisers for CYP2D6 substrates. Extensive studies focusing on the complex polymorphisms of *CYP2D6* have produced algorithms to predict the phenotypic response of patients on anti-depressant treatment (Gaedigk et al., 2008; Kirchheiner, 2008).

5.3.5. FMO3

Reports on the distribution of *FMO3* polymorphisms among populations are mainly confined to North America (Hao et al., 2007; Lattard et al., 2003). Due to the long immigration history and large admixture, the genetic backgrounds of the socially defined ethnic groups on that continent are not always as clear as they appear to be. It may therefore be misleading to extrapolate the results to corresponding ethnicities on other continents.

Here, the most diverse sample of African populations was analysed for variants in the *FMO3* gene. Only SNPs that are predicted to affect enzyme function were selected. The g.15167 G>A (E158K) variant was

detected at higher frequencies in Africans, comparable to Caucasians (Table 14), but all other SNPs were either found at low frequencies - g.18281 G>A (V257M), g.21443 A>G (E308G) and g.15089 G>C (D132H) – or absent - g.21599 T>C (L360P) - in the African populations. The G308/K158 compound variant also occurs at low frequencies, and its relevance in Africans is not clear, although it has implications in response to FMO3 substrates such as trimethylamine (Zschocke et al., 1999).

In summary, this data strongly confirms the rationale of the re-sequencing strategy of this study and indicates that this strategy should be equally applied to the *FMO3* gene in the future. As it appears that *FMO3* features a relatively low number of non-synonymous SNPs (Koukouritaki et al., 2007), genotyping and functional characterisation of variants might turn out to be easier than in other genes.

Variation in FMO3-dependent metabolic activity appears to be primarily associated with genetic variation (Koukouritaki and Hines, 2005). Pharmacogenetic applications based on FMOs are not yet realised, although there is much interest in determining the level of phenotypic variation across world populations (Mao et al., 2008). The ability of this enzyme to bypass the production of reactive oxygen species (as done by CYPs) makes it a favourable target for metabolism of drug substrates (Krueger and Williams, 2005), as many CYP substrates can also be metabolised by FMOs. Therefore, FMO3 may provide an important alternative route for drug elimination if the respective CYPs are inhibited or functionally impaired.

5.3.6. GST

The frequencies of *GST* homozygous deletions found here are comparable with previous studies (Cotton et al., 2000). Both *GSTM1* and *GSTT1* deletions are mostly less prevalent in Africans than in Asians (Table 14). Whereas *GSTM1* occurs as frequently in Caucasians as in Asians, the prevalence of *GSTT1* in Caucasians is only half of the Asian figures.

Individuals most seriously affected are those with a homozygous deletion of an entire gene. For example, *GSTM1* and/or *GSTT1* null alleles have been associated with colorectal cancer predisposition (Ates et al., 2005; Pande et al., 2008; Smits et al., 2003). In addition, these polymorphisms are risk factors for coronary artery disease in type 2 diabetic patients, particularly among smokers (Manfredi et al., 2009). Another disease risk associated with smoking and homozygous deletion of *GSTM1* and/or *GSTT1* is the development of asthma in adults (Saadat and Ansari-Lari, 2007). Therefore, individual genotyping for *GSTM1* and *GSTT1* could be informative for risk assessment if a person is exposed to certain environmental conditions such as smoking.

5.3.7. NAT2

Nineteen *NAT2* alleles have been discovered so far (Hein et al., 2008). The major defective alleles are *NAT2*5*, *NAT2*6*, *NAT2*7* and the African-specific *NAT2*14*. Individuals homozygous for those alleles, or being compound heterozygotes, can be predicted as slow acetylators. It has been speculated that the variation in acetylator status across major world populations reflects differences in dietary habits or the environment. There is a high prevalence of slow and intermediate acetylators in African populations, due to the *NAT2*5* (I114T), *NAT2*6*

(R197Q) and *NAT2*14* (R64Q) alleles. This is consistent with the *NAT2* allele frequency data of this study. Haplotype determination also concurs with *NAT2*5B* and *NAT2*6A* (Figure 4), being the most common in Africans, and with recent studies of *NAT2*4* in sub-Saharan populations (Sabbagh et al., 2008; Sabbagh and Darlu, 2006).

The influence of *NAT2* genotype on dosage and pharmacokinetics of isoniazid has been demonstrated in patients suffering from pulmonary tuberculosis, whereby slow acetylators require half the dose compared to fast acetylators to achieve satisfactory bactericidal activity (Donald et al., 2007). Cotrimoxazole remains the drug of choice in the prophylaxis and treatment of *Pneumocystis carinii* pneumonia in patients infected with HIV. This drug is metabolised by NAT2, and individuals carrying reduced function alleles may be exposed to higher amounts, which would be available for oxidative metabolism by CYP2C9, making them more susceptible to hypersensitivity reactions (Carr et al., 1994). In conjunction with environmental toxins and polymorphisms in other genes, such as *GSTM1* and *GSTT1*, *NAT2* may contribute to bladder and breast cancer risk (Lee et al., 2003; McGrath et al., 2006). Therefore, *NAT2* genotypes are of clinical relevance in Africans for dosing of drugs used widely in treatment of TB and HIV patients as well as for determining susceptibility to disease.

5.4. Population differentiation and evolutionary relatedness

Allele frequency data of DME genes has been used to examine the extent of differentiation and relatedness of African ethnic groups compared to other world populations (Figure 5). Statistical significance of allele frequency differences was gauged with Fischer's exact test and F statistic. Correlations of genetic and geographic distance were evaluated under an isolation by distance model. PCA and phylogenetic UPGMA analysis were applied to determine population relatedness. Eventually, the hierarchical structure of population variance was assessed by AMOVA.

Earlier studies of population relatedness have usually focused on microsatellite, mitochondrial and Y chromosome markers, beside gene families such as *HLA* (Agrawal et al., 2007; Li et al., 2008). Illustrations of genetic diversity in the geographical space, with allele frequency distributions and concordant clustering of populations, were generated, showing geographical/continental structures (Novembre et al., 2008; Reich et al., 2008). However, pharmacogenetic markers are still underrepresented in such studies (Sistonen et al., 2009), and the extent to which ethno-linguistic classifications may reflect pharmacogenomic diversity is not clear.

5.4.1. Differentiation of African populations

World sample collections such as the Human Genome Diversity Panel from the Centre d'Etude du Polymorphisme Humain (Cann et al., 2002) only contain small sets of African populations. In some cases, ethnic groups available in public databases such as Yoruba of Ibadan (YRI) (www.hapmap.org) or African Americans are supposed to

represent all Africans, an assumption that is largely insufficient, considering the high level of genetic diversity in African populations. Consequently, in this study, African populations were selected to represent at least four of the major ethno-linguistic groups and to cover three of the five main geographic parts of Africa (Table 1). It is assumed that such a selection should be suitable for investigating population differentiation and evolutionary relatedness of sub-Saharan African populations.

Re-sequencing data was analysed for genetic differentiation at each gene locus. Genotype frequency differences amongst populations were statistically significant at *CYP2C19*, *CYP2D6* and *NAT2*, but not at *CYP2C9* (Tables 8-11). This could be due to a more homogenous nature of the *CYP2C9* locus, compounded by the relatively low frequency of SNPs detected in this gene. However, average Fst values show low levels of population differentiation for all genes, as would be expected when analysing populations from the same continental region. A study on world populations of over 1,000 microsatellite markers suggested that geographic distances may be a better predictor of genetic diversity than ethno-linguistic classifications (Belle and Barbujani, 2007). With this in mind, the relevance of geographical distance for diversity of African genes was assessed using an isolation by distance model (see Table 15). Low values of correlation were obtained for all genes, with only *CYP2C9* proving statistically significant and *CYP2D6* showing no correlation at all. Whereas the *CYP2C9* result might be influenced by the low frequency of SNPs (see above), the *CYP2D6* data, compounded by the negative weighted Fst value, would be consistent with this gene's highly polymorphic character, featuring several altered activity variants that are not

showing a conclusive geographical pattern either (Sistonen et al., 2007).

Frequency data from the HapMap SNP analysis was used to determine the extent of population differentiation among Hausa, Maasai, San and Shona, compared to Yoruba (YRI). It appears that *NAT2* SNPs present the highest significance for differentiation of these populations, suggesting a comparatively more heterogeneous nature of this gene in Africans.

Only *NAT2* SNP frequencies appear to significantly vary among the four populations ($P<0.05$), with rs1801280, rs4646246, rs7832071 showing the highest Fst values (see Table 13). Such variability of Fst values for SNPs across the *NAT2* gene region has been attributed to population-specific selective pressures acting on the various DME loci (Sabbagh et al., 2008). When *NAT2* data were removed from the analysis, less differentiation between population pairs was found (see Table 16B).

Phylogenetic analysis of populations, using UPGMA and taking the frequencies of all thirty SNPs into account, showed clearly that Maasai separates from the rest of the populations and to some extent, the YRI forms its own branch (see Figure 6). These findings further indicate the limitations of the use of HapMap data for inferring SNP frequencies for supposedly closely related populations, since some gene regions may be more heterogeneous than others and variable in some populations more than in others.

5.4.2. Evolutionary relatedness based on common alleles

Allele frequencies of DME genes in African, Asian and Caucasian populations were used to explore evolutionary relatedness by Principal Component Analysis (PCA, see Fig 7). The results clearly show distinct clustering of the main world population clusters, in agreement with studies on other genes (Cavalli-Sforza, 2005). Projecting genetic variation onto an evolutionary tree, it was proposed that certain alleles are very old and may pre-date the splitting of the three major world populations (Aklillu et al., 2007). For example, *CYP2D6*5* is found at similar (low) frequencies in all populations (see Table 14). In contrast, alleles such as *CYP2D6*17*, *CYP2D6*29* and *NAT2*14* are African-specific and seem to have occurred after the departure of populations that later formed the Caucasian and Asian groups. However, alleles such as *CYP2D6*4* and *CYP2D6*10*, most prevalent in Caucasians and Asians, respectively, are both present at low frequencies in other populations too, indicating that they occurred before the departure of those populations but were selected on differently under different environmental conditions, as opposed to earlier suggestions. It could also be that their presence in the other populations is due to population admixture. For example, African Americans are said to have up to 20% of their genetic make-up derived from Europeans (Halder et al., 2008; Parra et al., 1998; Shriver et al., 2003), hence the occurrence of *CYP2C9*2* in African Americans at a frequency of 3%, compared to its absence in sub-Saharan African populations, where lower levels of admixture might have occurred.

UPGMA results show the same separation of the major world population groups (see Figure 8A). Analysing Africans alone reveals a pattern that is difficult to explain in terms of geography or ethnicity

(see Figure 8B). However, it shows all the groups of the Bantu sub-family clustered on one branch side. Surprisingly, Tanzanian Bantu cluster with Luo, who are Nilotic, although they would be expected to be closer to the Maasai. The west African populations appear on the other side of the cluster, whereas African Americans and Ghanaians are in the middle. This is in agreement with African Americans being an admixture of mostly western and southern African populations (71%), belonging to the Niger Congo family from the south (Tishkoff et al., 2009).

Interestingly, San and Ethiopians are distinctly separated from the rest of the African populations. The San separation is not surprising, since microsatellite and mitochondrial analyses have revealed the same pattern (Tishkoff et al., 2009). Ethiopians are from the Afro-Asiatic ethno-linguistic family, however mostly Semitic and Cushitic subfamily, which is separate from the Hausa, who belong to the same family but are Chadic. The Ethiopian separation may also be exaggerated by the high prevalence of the multiplication alleles *CYP2D6*2XN*, compared to other African populations.

Building upon the significant differences in allele frequencies (see Table 17), the PCA (see Figure 7) and UPGMA analyses (see Figure 8), AMOVA was used to distinguish how different levels of population groups contribute to the total variation observed (see Table 18). The AMOVA analysis showed clearly that most of the global variance occurs **within** populations or ethnic groups, (>=97%). In other words, variance is pretty evenly spread out over world populations, with the differences **among** them being small by comparison. The effect becomes even more pronounced within African ethnic groups arranged

by country, geographical region or ethno-linguistic family, with variation running up to 99.75%, 99.82% and 99.91%, respectively. This confirms earlier reports, finding 97-98% variance of microsatellite markers within African ethnic groups (Chen et al., 2005). In summary, this data indicates that genetic diversity **within rather than among** population groups is the main driver of personalisation.

5.4.3. Pharmacogenetic variants as population markers

The results of this study support the concept of using genetic markers for inferring evolutionary relations between populations, particularly at the world and continental level (Belle and Barbujani, 2007; Tishkoff et al., 2009). However, very few earlier studies have used gene markers directly associated with phenotypic variation (Sabbagh et al., 2008; Sistonen et al., 2009). Instead, microsatellite markers have been deemed more reliable for assessing genetic diversity (Belle and Barbujani, 2007), due to their neutral nature, in contrast to SNPs in coding genes that may be under different selective pressures, suffer from ascertainment bias and hence may not have enough power to furnish a detailed population structure. In addition, it was suggested that mitochondrial markers provide clearer and more conclusive patterns for refining ethno-linguistic sub-families (Behar et al., 2007; Behar et al., 2008). However, the genome is a heterogenous landscape, and variation is not restricted to microsatellite and mitochondrial markers, which are both equally insufficient in explaining phenotypic diversity.

Although microsatellite markers seem sufficient for anthropological purposes, scanning geographical distances and ethno-linguistic history, they do not offer information on phenotypic variation relevant to medical applications. Therefore, this study makes a timely contribution

as it follows up on previous suggestions to characterise functional variants in Africans associated with disease and drug response (Tishkoff et al., 2009), in particular towards understanding pharmacogenetic diversity. Here, it was confirmed that functional variants can be used as markers to separate the world populations (Africans, Asians and Caucasians) and infer some degree of differentiation between more closely related ethnic groups. In fact, the pattern obtained with pharmacogenetic markers did not differ considerably from earlier reports with larger marker sets and more diverse population samples (Tishkoff et al., 2009).

Nevertheless, the observation that the highest level of variation is found between individuals within populations suggests that a larger number of markers could enable refined separations, producing a more detailed population sub-structure of African ethnic groups. Such endeavour would require localised sampling approaches with well ascertained variants, based on future expansion of African population resources (see below). However, neither a larger number of samples nor increasing the set of markers guarantee a better description of genome diversity, compounded by the high prevalence of low frequency and rare alleles. With this in mind, the current use of high frequency alleles is regarded a start towards building phenotype clusters in Africans with a determinant set of pharmacogenetic markers.

5.5. Africa: Bioresource building

Africa displays vast geographic, linguistic and cultural diversity (Sirugo et al., 2008). In addition, African genetic history supports a high level of variation in the genes of African people (Tishkoff et al., 2009). Yet this diversity is still totally detached from medical practice in Africa. Pharmacogenetics is an important tool to change this and translate genetic information into improved disease diagnosis and treatment, targeted specifically at African populations. Unfortunately, due to limited funding, technical capacity and resources, African pharmacogenetics is still in its infancy. The establishment of bioresources, such as biobanks and genetic databases, is an essential part of the effort to translate this research into clinical practice (Matimba et al., 2008).

The impact of pharmacogenetics on targeted therapy is well demonstrated and starting to be implemented in the developed world (Rahemtulla and Bhopal, 2005) (see Table 3). Yet pharmacogenetic studies in African populations are still few and scattered, with little clinical use so far (Table 22). It was the aim of this study to contribute to this growing field by building pharmacogenetic resources in Africa and thereby improving the representation of Africans in drug discovery and development, facilitating personalised medicine in Africa (see Figure 9).

5.5.1. African populations

The sample collection of this study spans three geographic regions in sub-Saharan Africa. Based on demographic information, this diverse collection would be representative of at least 150 million Africans. Ethnicities of the Bantu subfamily (Niger Congo B) constitute the

largest group in sub-Saharan Africa, stretching from west, central, across to east and largely populating southern Africa (Vansina, 1995). Here, Kikuyu, Shona, Tanzanian Bantu and Venda represented this subfamily. Yoruba and Igbo also belong to the Niger-Congo A subfamily and would represent populations in west Africa, including Benin, Ghana, Ivory Coast, Nigeria and Senegal. In addition, most African Americans originated from west Africa and therefore, should also be represented by these populations. Hausa belong to the Afro-Asiatic family group, mainly found in the northern part of Africa. Nilo-Saharan Luo and Maasai represent ethnic groups in eastern African countries such as Kenya and Tanzania, including Sudan and Uganda.

Ethno-linguistic classifications according to Ethnologue classifications (Gordon, 2005) were used to assign population groups (see Table 1). However, most ethnic groups are variants or dialects of the four major African language classes, thus exaggerating the number of population clusters as seen in some countries, such as Nigeria or Tanzania, where there are over 200-300 ethnic groups. Ethnic groups in close proximity are likely to have less differences and belong to the same ethno-linguistic subfamily. On the other hand, clustering subjects by country does not accurately reflect ethnicity either. Due to the economically-driven partition of Africa in the 1800's, people of different 'ethnicity' were lumped together or those of the same group split between two or three nations. Therefore, there is a tendency to overestimate the variation among ethnicities of the same subfamily. Accordingly, AMOVA analysis (see Table 18) shows that little variation occurs among ethnic groups, suggesting that individual variation plays the most important role in explaining phenotypic diversity based on pharmacogenetics markers.

Here, samples from five African countries (Kenya, Nigeria, South Africa, Tanzania and Zimbabwe) were collected, using self-proclaimed ancestry. However, admixture cannot be ruled out, since claims were based on parents' and grandparents' belief of which ethnic group they belonged to. To ascertain the true ancestry of samples, parallel analysis of mitochondrial and Y-chromosome markers would be advisable. In cases were university students were chosen as representative for ethnic groups, bias of area of origin cannot be ruled out. However, with challenges of sample collection using 'random' methods, university students represent individuals from different areas/villages in the particular country and therefore may well represent sufficient geographical background diversity.

Medical research and ethics review boards are at different levels of development in African countries, ranging from non-existent to those reviewing old guidelines to integrate recent advances in genomics research. This situation had implications for the criteria of this study, limiting them to pharmacogenetics of drug metabolising enzymes. To comply with minimum requirements, a sample repository was set up, to ensure that enough genetic material was available. All samples were anonymised, reducing the sensitivity of sample/result confidentiality. This was considered sufficient, while more rigorous laws about sample collection activities are being drawn out and permanently revised to fit regional and international recommendations currently under review in Africa.

5.5.2. African pharmacogenetics

The status of pharmacogenetics research in Africa was assessed by a literature scan via NCBI PubMed (www.pubmed.gov), producing a total of 52 publications focused on drug metabolising enzyme genes in

African populations over the last 15 years (Table 22). Information was recorded according to the following criteria: Population or ethnic group, country of study, city or location in the respective country, genes and alleles included, type of study (genotyping and/or phenotyping), drugs or substrates under investigation, on either healthy volunteers or disease patients, and finally methods employed such as RFLP, re-sequencing and others.

Overall, the number of publications on African pharmacogenetics increased in recent years, but remained largely restricted to drug metabolising enzymes. The most studied genes in African populations are CYPs, followed by GST, NAT and TPMT. Since most analyses were based on findings in Caucasians, many of the analysed polymorphisms may not be relevant in Africans, adding to the already critical question of ascertainment bias.

It appears that pharmacogenetics studies have been reported on populations of at least 17 African countries (Table 22), mostly in eastern and southern Africa (Figure 13), leaving a gap for the extension of such work to other parts of the continent.

Methods employed for characterisation of polymorphisms were mainly PCR-RFLP at lower throughput, with most sample sizes limited to a maximum of a few hundred. Genotype-phenotype correlations are included in a minority of reports, mostly focused on healthy volunteers. The most common clinical associations include HIV patients on anti-retroviral treatment such as efavirenz and nevirapine, linked to the respectively most relevant polymorphism *CYP2B6*6*. Most studies involve *CYP2C19* and *CYP2D6*, including some clinically important drugs that are already on the market. There are no pharmacogenetic

reports on drug substrates in development or as part of a retrospective African clinical trial. Considering the advances on pharmacogenetic markers as shown in Table 3, it appears that African studies are small scale and less focused on progress towards personalised solutions. In summary, more extensive analyses are required, with larger samples sizes, to ascertain for as many polymorphisms as possible, which may contribute to phenotypic variation.

Table 22: Pharmacogenetics studies in African populations

Population, source/location[1]	Gene(s)	Study type	Drugs/substrates	Subjects	No.	Methods	Reference
African, HGDP	*NAT1, NAT2*	G, A		Healthy volunteers	-	Sequencing	Patin et al., 2006
African, HGDP	*CYP2D6*	G, A		Healthy volunteers	-	LR-PCR primer extension	Sistonen et al., 2007
Angolan, Cabinda, Angola	*TPMT*	G		Healthy volunteers	103	HCSGE	Oliveira, 2007
Beninese	*CYP2C9, CYP2C19*	G		Healthy volunteers	111	AS-PCR	Allabi et al., 2003
Beninese	*CYP2C9*	G, P	losartan	Healthy volunteers	19	RFLP	Allabi et al., 2004
Beninese	*CYP2C9, CYP2C19*	G, P	phenytoin	Healthy volunteers	109	RFLP, Sequencing	Allabi et al., 2005
Burkina Faso	*CYP2C8*	G,P,C	amodiaquine	Children malaria patients	275	RT-PCR	Parikh et al., 2007
Egyptian, Cairo	*ABCB1*	G, P	phenytoin	Epileptic patients and controls	150	RFLP	Ebid et al., 2007
Egyptian, Cairo	*CYP2C9, CYP2C19, CYP2E1, DPYD*	G		Healthy volunteers	247	AS-PCR	Hamdy et al., 2002
Egyptian, Cairo	*CYP1A2, GSTM1, GSTT1, SULT1A1, TPMT*	G		Healthy volunteers	212	RFLP	Hamdy et al., 2003
Ethiopian, Addis Ababa	*CYP2D6*	G, P	debrisoquine	Healthy volunteers	122	LR-PCR, RFLP	Aklillu et al., 1996
Ethiopian, Addis Ababa	*CYP1A2*	G,P	caffeine	Healthy Volunteers	100	RT-PCR, Sequencing	Aklillu et al., 2003
Ethiopian, Addis Ababa	*CYP2D6*	G, P	debrisoquine	Healthy volunteers	115	LR-PCR, RFLP	Aklillu, 1996
Ethiopian, Addis Ababa	*CYP2C19*	G,P	S-mephenytoin	Healthy Volunteers	114	RFLP	Persson et al., 1996
Ethiopian, Addis Ababa	*CYP2C9*	G		Healthy volunteers	150	RFLP	Scordo et al., 2001
Ethiopian, Addis Ababa	*CYP2C9*	G		Healthy volunteers	150	RFLP	Yasar et al., 2002
Ethiopian, Stockholm	*CYP2C19*	G,P	omeprazole	Healthy Volunteers	126	Sequencing	Sim et al., 2006
Gabonese	*CYP2D6*	G, P	dextromethorpan	Healthy Volunteers	154	LR-PCR, RFLP	Panserat et al., 1999
Gambian	*GSTM1, GSTT1, GSTP1*	G,P	environmental factors	HBV patients	357	PCR	Wild et al., 2000
Ghanaian, Accra	*CYP2D6*	G, P	dextromethorphan, debrisoquine, sparteine	Healthy volunteers	21	LR-PCR, RFLP	Droll et al., 1998
Ghanaian, Accra	*CYP2D6*	G, P	debrisoquine, sparteine	Healthy volunteers	326	LR-PCR, RFLP	Griese et al., 1999
Ghanaian, Accra	*CYP2B6*	G		Healthy volunteers	64	Sequencing	Klein et al., 2005
Ghanaian, Accra	*CYP2B6*	G		Healthy volunteers	42	LDR-FMA	Mehlotra et al., 2006
Ghanaian, Accra	*CYP2B6*	G		Healthy volunteers	33	LDR-FMA	Mehlotra et al., 2007
Ghanaian	*CYP3A4*	G		Healthy volunteers	100	SSCP	Tayeb et al., 2000
Ghanaian	*TPMT*	G		Healthy volunteers	217	AS-PCR, RFLP	Ameyaw et al., 1999
Northern Ghana	*CYP2C8*	G,P	amodiaquine	Children malaria patients	200	RFLP	Rower et al., 2005
Guinean, Guinea Bissau	*CYP2B6*	G		Healthy volunteers	21	LDR-FMA	Mehlotra et al., 2006
Guinean, Guinea Bissau	*CYP2B6*	G		Healthy volunteers	32	LDR-FMA	Mehlotra et al., 2007
Ivorian, Ivory Coast	*CYP2B6*	G		Healthy volunteers	41	LDR-FMA	Mehlotra et al., 2006
Ivorian, Ivory Coast	*CYP2B6*	G		Healthy volunteers	45	LDR-FMA	Mehlotra et al., 2007
Kenyan, Nairobi	*TPMT*	G		Healthy volunteers	101	RFLP	McLeod et al., 1999
Mozambican, East Coast	*TPMT*	G		Healthy volunteers	250	HCSGE	Alves, 2004
Senegalese, Senegal	*CYP2B6*	G		Healthy volunteers	10	LDR-FMA	Mehlotra et al., 2006
Sierra Leonian	*CYP2B6*	G		Healthy volunteers	52	LDR-FMA	Mehlotra et al., 2006

Table 22: Pharmacogenetics studies in African populations (continued)

Population, source/location[1]	Gene(s)	Study type	Drugs/substrates	Subjects	No.	Methods	Reference
Venda, South Africa	*CYP2C19, CYP2D6*	G		Healthy individuals/ psychiatric patients	76	RFLP	Dandara et al., 2001
Venda, South Africa	*CYP1A1, GSTM1, GSTT1, GSTP1*	G		Healthy Volunteers	76	PCR, RFLP	Dandara et al., 2002
Venda, South Africa	*NAT2*	G		Healthy Volunteers	96	RFLP	Dandara et al., 2003
South African	*CYP3A4, ABCB1*	G		Healthy volunteers	220	RFLP	Chelule et al., 2003
South African	*CYP2E1*	G		Healthy volunteers	331	RFLP	Chelule et al., 2006
South African	*NAT2*	G,P	isoniazid	PTB patients	66	RFLP	Donald et al., 2007
South African	*CYP3A5, CYP2B6, CYP3A4*	G,P,C	nevirapine	HIV patients	385	RT-PCR	Haas et al., 2006
South African	*TPMT*	G,P	azathioprine	rheumatology patients	465	RFLP	Heckmann et al., 2005
Northern Sudanese	*NAT2*	G		Unrelated	127	AS-PCR, RFLP	Al-Yahyaee et al., 2007
Bantu, Tanzania	*CYP2C19*	G,P	mephenytoin, proguanil	Healthy Volunteers	195	RFLP	Bathum et al., 1999
Bantu, Tanzania	*CYP2D6*	G,P	sparteine	Healthy Volunteers	196	LR-PCR, RFLP	Bathum et al., 1999
Bantu, Tanzania	*CYP2D6*	G		Healthy individuals/ psychiatric patients	194	LR-PCR, RFLP	Dandara et al., 2001
Bantu, Tanzania	*CYP1A1, GSTM1, GSTT1, GSTP1*	G		Healthy Volunteers	100	RFLP	Dandara et al., 2002
Bantu, Tanzania	*NAT2*	G		Healthy Volunteers	117	RFLP	Dandara et al., 2003
Bantu, Tanzania	*CYP1A2*	G		Healthy Volunteers	71	RFLP	Dandara et al., 2004
Bantu, Tanzania	*CYP2D6*	G,P	debrisoquine	Healthy volunteers	106	LR-PCR, RFLP	Wennerholm et al., 1999
Tanzanian, Unguja and Pemba, Zanzibar	*CYP2B6, CYP3A4, CYP3A5*	G		Malaria patients	103	RFLP, RT-PCR	Ferreira et al., 2008
Zanzibar	*CYP2C8*	G,P	amodiaquine	malaria patients	165	RFLP	Cavaco et al., 2005
Tanzanian, Dar es Salaam	*CYP2C19*	G,P	mephenytoin, omeprazole	Healthy volunteers	251	RFLP	Herrlin, 1998
Tanzanian, Dar es Salaam	*CYP2C9*	G		Healthy volunteers	183	RFLP	Yasar et al., 2002
Ugandan, Kampala	*CYP2B6, CYP3A4, CYP3A5*	G,P	nevirapine	HIV patients	23	RFLP	Penzak et al., 2007
Shona, Zimbabwe	*CYP2C19, CYP2D6*	G		Healthy individuals/ psychiatric patients	114	RFLP	Dandara et al., 2001
Shona, Zimbabwe	*CYP1A1, GSTM1, GSTT1, GSTP1*	G		Healthy Volunteers	100	RFLP	Dandara et al., 2002
Shona, Zimbabwe	*NAT2*	G		Healthy Volunteers	163	RFLP	Dandara et al., 2003
Shona, Zimbabwe	*CYP1A2*	G		Healthy Volunteers	143	RFLP	Dandara et al., 2004
Shona, Zimbabwe	*CYP2C19*	G,P	mephenytoin	Healthy Volunteers	103	RFLP	Masimirembwa et al., 1995
Shona, Zimbabwe	*CYP2D6*	G,P	debrisoquine, metropolol	Healthy volunteers	103	LR-PCR, RFLP	Masimirembwa et al., 1996
Shona, Zimbabwe	*GSTM1*	G		Healthy Volunteers	148	PCR	Masimirembwa et al., 1998
Shona, Zimbabwe	*GSTT1*	G		Healthy Volunteers	123	PCR	Masimirembwa et al., 1998
Zimbabweans, Harare	*CYP2B6*	G,P	efavirenz	HIV patients	70	RFLP	Nyakutira et al., 2008

[1]Source of samples, if part of a collection (e.g. HGDP) or geographical location/city (if known); Populations are identified by country name, e.g. Beninese=from Benin. Country is indicated if ethnic group name is used; No.=number of samples in study; HGDP=Human Genome Diversity Panel; PCR=polymerase chain reaction; AS=allele-specific; RT=real time, RFLP=restriction fragment length polymorphism; LR=long-range, SSCP=single strand conformation polymorphism; LDR-FMA= multiplex ligase detection reaction-fluorescent microsphere assay; HCSGE= horizontal conformational sensitive gel electrophoresis; DPYD= dihydropyrimidine dehydrogenase; SULT=sulfotransferase; PTB=pulmonary tuberculosis; G=genotyping, P=phenotying, A=anthropology, C=clinical trial.

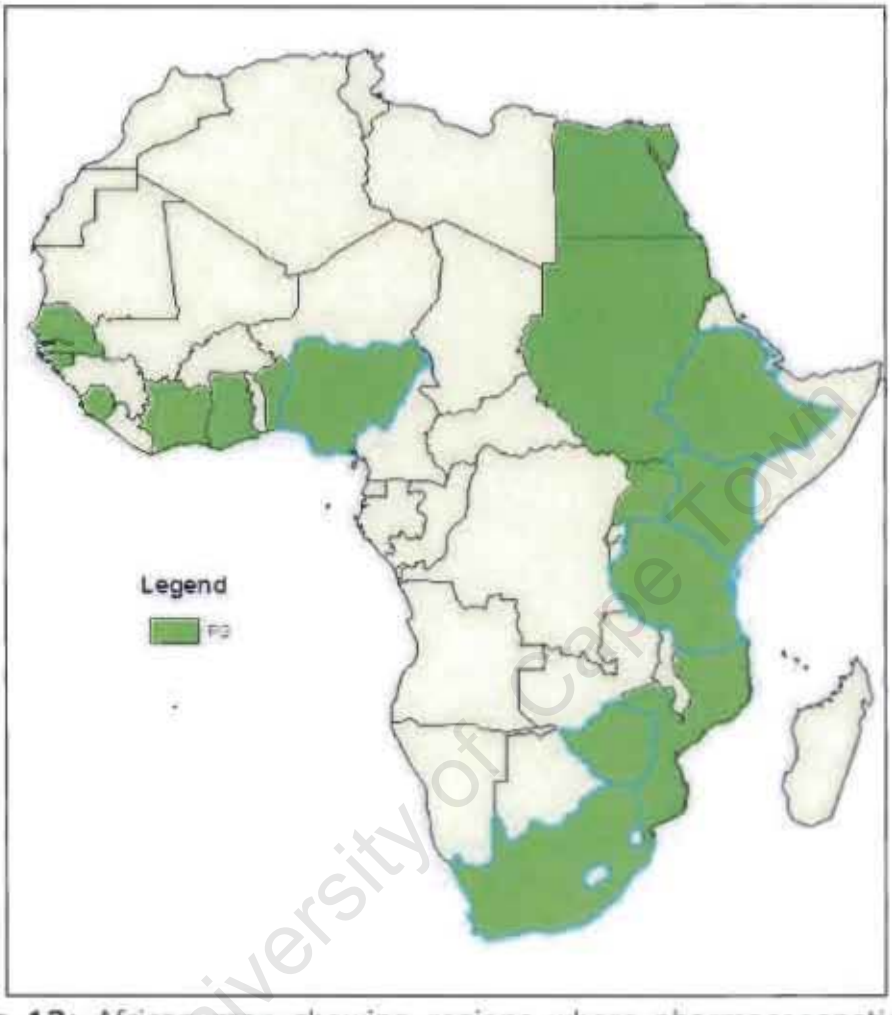


Figure 13: African map showing regions where pharmacogenetics studies (PG) have been carried out. Highlighted are the countries from where ethnic groups have been sampled for characterisation of NAT2 genes. This figure was generated from Geographical Information Systems and Africa Genetics project (Matimba et al., unpublished data).

5.5.3. African biorepositories

Global advances in biobanking are synonymous with progress towards personalised healthcare. Whereas this concept is taking root in many places now (see 1.3.2.2), fostering focused, local solutions of translational medicine, Africa lags behind with only a few biobanks established so far. Sample collections of African people include a DNA

bank in the Gambia (Sirugo et al., 2004), a proposed African Biobank in Cameroon (Jackson, 2006) and the African American Population Biobank in the USA (Kaiser, 2003). These and other small-scale collections in research institutions are restricted to specific regions or countries and may not capture the whole range of ethnic diversity in Africans.

Therefore, coordinated efforts are required to build facilities and establish collaborative programmes in the region. A more focused approach is necessary to harvest the benefits of the latest advances in genetics, disease diagnostics, drug discovery and development. To increase statistical power of pharmacogenetics studies, more large scale sample collections, socio-demographic data and clinical information from volunteers and patients must be enabled, while facilitating access to these resources for researchers and pharmaceutical companies.

As more than one thousand samples were collected for this study, it may be appropriate that this material be extended to efforts of building such a bioresource. Consequently, the establishment of a Biobank of African Populations has been proposed (Matimba et al., 2008).

5.5.4. African Databases

5.5.4.1. Sample and genotype database

A database recording individual sample information and genotype data was set up, thereby emulating an information management system for centralised control of sample analysis and experimental procedures (see Figure 10). This enabled the development of additional search

criteria to call up genotype frequencies and graphical comparisons among populations (see Figure 11). The resource is of primary importance for routine use in handling large sample collections for genetic and diagnostic analysis and should be incorporated into an integrated pharmacogenetics database system.

5.5.4.2. Pharmacogenetics database

Based on the catalogue of African polymorphisms in drug metabolising enzyme genes, together with the extensive sample collection from this study, an allele frequency database has been proposed, following earlier efforts towards the establishment of a pharmacogenetics database of African populations (Matimba et al., 2008). So far, allele frequency data of a total of 180 polymorphisms in 10 populations have been recorded (see Tables 19 and 20). In addition, polymorphisms from other African populations were included in this catalogue, creating baseline information on the prevalence of polymorphisms in these populations. The immediate use of this information for determining variation among Africans and other world population groups is illustrated in Table 17, Figures 7 and 8.

On a global scale, research networks, such as NIH Pharmacogenetics Research Network (PGRN), and associated databases, such as the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB), became valuable resources in the field (Giacomini et al., 2007; Gong et al., 2008). Using them, pharmacogenes of importance for the metabolism of prescribed drugs and detoxification of xenobiotics have been selected (Sangkuhl et al., 2008). Now referred to as VIP (very important pharmacogenes), these include thirty drug receptors, transporters and drug metabolising enzymes, about a quarter of which are CYPs. Four CYP genes were characterised here,

encompassing most of the polymorphisms in coding regions reported in Africans to date. To complement this dataset, other polymorphic genes of importance for the metabolism of pharmaceutical drugs and environmental toxins were included: *FMO3*, *GSTM1*, *GSTT1* and *NAT2* (see Table 4). So far, most African studies had been confined to genotyping, with few re-sequencing analyses, hence limiting the ascertainment of African polymorphisms and haplotypes. Therefore, this study adds considerably to the ever growing record of variants and the global pharmacogenetic knowledge base.

Summary information on pharmacogenetic studies in Africa has been collected, focusing on major drug metabolising enzymes (Table 22). This resource can be used for mapping pharmacogenetic knowledge (Figure 13) and gauging the level of understanding, in order to find out where critical information is lacking. For example, in contrast to extensive studies on the effects of *CYP2D6* on metabolism of drugs such as debrisoquine and dextromethorphan in Africans (Table 21), comparable analyses on the impact of *CYP2C9* polymorphisms on warfarin metabolism are scarce in Africans. With the help of a pharmacogenetics database of African populations, the prevalence of altered activity variants, associated with a known clinical phenotype in other populations, could be quickly established during therapy or clinical trials.

In conclusion, this study represents the first successful consolidation of polymorphisms in drug metabolising enzyme genes in African ethnic groups towards the curation of a knowledge base and development of an information portal of African pharmacogenetics.

5.6. PGxA - Pharmacogenetics for Personalised Medicine in Africa

Pharmacogenetics is firmly on its way into the doctor's office in order to find safer and more effective drug regimens (Grossman, 2007; Roses et al., 2007). Based on patient clinical data as well as population variation, the FDA has started to modify labels of some medicines to include pharmacogenetic information (see Table 3).

Africa, with its specific disease and genetic landscape, needs to develop pharmacogenetic research and its clinical applications in order to reap their benefits for its medical practice.

5.6.1. Benefits for treatment and exposure

Polymorphisms in drug metabolising enzyme genes affect the pharmacokinetics of their drug substrates, causing toxicity and resistance due to suboptimal therapeutic levels (Nolan et al., 2006; Rotger et al., 2005). Due to the disease burden bias, anti-infectives such as anti-HIV, anti-malaria and anti-tuberculosis drugs are widely used in Africa. Most of these drugs induce adverse reactions related to their pharmacokinetics, have narrow therapeutic indices, dose-limiting toxicity and large inter-individual variation in plasma concentration. Therefore, the clinical use of such drugs should benefit from pharmacogenetic information.

5.6.1.1. Minimising ADRs

Drugs optimised for use in Europe may need to undergo specialised assessments for use in Africans. For example, it has been observed that Africans suffer more severe side effects due to anti-retroviral drugs such as efavirenz and nevirapine, which are metabolised by

CYP2B6. This is due to the *CYP2B6*6* allele, which causes impaired metabolism of these drugs (Rotger et al., 2005d) and, as shown in this study, occurs at high frequencies in Africans. Hence it was proposed that there may be a need for dosage adjustment in adults on efavirenz treatment (Nyakutira et al., 2008). Furthermore, other alleles such as *CYP2B6*18*, which was recently reported in Africans (Mehlotra et al., 2007; Wang et al., 2006), may contribute to inter-individual variability, but studies are still limited to a few African populations.

5.6.1.2. Improving drug economy

In situations where it can be demonstrated that individuals carrying reduced activity variants would be effectively treated by using half the dose (Nyakutira et al., 2008), as in the above case of efavirenz, such treatment can be made available to more patients at a lower cost. Since drug therapy is expensive in Africa, this would go a long way in addressing economic limitations in countries where governments are offering large scale treatment schemes.

5.6.1.3. Monitoring patient compliance

Side effects from drug regimes such as HAART (highly active anti-retroviral treatment) may cause patients to discontinue their treatment schedule or skip doses due to feeling poorly (personal communication). This poses a danger of sub-therapeutic levels triggering the selection of resistant viral populations. Pharmacogenetic testing in conjunction with therapeutic drug monitoring may help to monitor patient compliance and design alternative treatment follow-up procedures.

5.6.1.4. Responding to environmental toxins

In most parts of Africa, poor storage of grain has caused food contamination by fungus, producing toxins such as Aflatoxin B1, which has been implicated in increased susceptibility to cancer (Slone et al., 1995; Wojnowski et al., 2004). The role of DMEs and drug transporters in bio-activation, detoxification, and disposition of these xenobiotics is a major point for determining how individuals or ethnic groups are going to respond upon exposure. Genetic association between DMEs and the development of cancer has been suggested in a few studies on African populations (Chen et al., 1996; Dandara et al., 2005; London et al., 1995). The level of genetic variation, as detected in this study, makes these genes useful tools for assessing the impact of genetic polymorphism on Africans' response to environmental toxins.

5.6.2. Pharmacogenetics for personalised therapy

Inter-ethnic differences in drug disposition have been shown to warrant population-specific prescription. For example, after finding that African Americans responded better to the drug Bidil, with less side effects than in Caucasians, this medicine is now being marketed specifically for African Americans (Kahn, 2008), suggesting that ethnicity may play an important role in drug prescription. Although the genetic contribution is not clear, this is the first and most successful example of this role in African pharmacogenetics so far, due to advanced applications of clinical studies in the USA (see 1.4.5).

In order to advance medical application of pharmacogenetics in Africa, an assessment of the status quo was initiated here (see Table 22). It turns out that there is still little research, concentrated in a few places.

Most studies are based on small sample sizes of patients or volunteers and lack phenotype information. This has kept the number of pharmacogenetic markers small, mainly confined to the usual CYP genes.

To overcome these limitations, a more localised approach is needed, whereby specific populations are sampled to validate pharmacogenetic markers relevant to them. Such information will ultimately be used to target sub-populations for which a drug may be more effective and have less ADRs and help to understand why some drugs fail in early clinical trials. Drugs tested on Caucasian populations may require efficacy/safety assessment and/or dosage adjustment in Africans. In reverse, some drugs that fail in Caucasians may be useful in Africans (Kahn, 2008).

In order to realise sampling of specific populations for personalised therapy in Africa, coordinated efforts in bioresource building and technical capacity development are required (see 1.5). In this study, the development of a biobank network of African populations is proposed as a follow-up to a previous study (Matimba et al., 2008). Based on existing pharmacogenetics research networks (Gong et al., 2008), additional sample collections should be added to create centralised virtual resources. Furthermore, a pharmacogenetics database (see 5.5.4.2) is proposed as an information portal, recording data from on-going pharmacogenetics studies in Africa. As an essential prerequisite for the clinical use of these resources and their associated genotype data, phenotype information needs to be collected (Snyder, 2009). For example, information on ADRs should be documented in a controlled manner, allowing researchers, clinicians and patients to

understand issues related to their treatment. By building these bioresources and networks, linking research and clinical applications, pharmacogenetics-informed personalised medicine may be on its way to becoming a reality in Africa.

The polymorphic genes of drug metabolising enzymes are the most important pharmacogenomic markers to date. Their potential impact on drug therapy is most clearly demonstrated for oral anticoagulants, such as warfarin and aceunocoumarol, which are so far the most successful examples of pharmacogenetic applications. One of the genes implicated in slow metabolism of these drugs is *CYP2C9*, with the most commonly associated polymorphisms being *CYP2C9*2* and *CYP2C9*3*. Interestingly, these alleles are rare in Africans, suggesting that other markers such as *CYP2C9*5*, *CYP2C9*6* and *CYP2C9*8* may be involved in inter-individual variation of these drugs' metabolism, as proposed previously (Matimba et al., 2009). In conclusion, genotype assessment is advised, including SNPs newly identified here, for optimising clinical drug use in African populations. This should enable correct dosage adjustment for individuals who are likely to experience ADRs owing to slow metabolism or an inadequate therapeutic effect caused by ultra-rapid metabolism.

5.6.3. A pharmacodiagnostic kit for African populations

Pharmacodiagnostic products for world populations are already on the market. The first FDA-cleared microarray for in vitro diagnostic use in the USA, the AmpliChip CYP450 Test (Roche Diagnostics, Indianapolis, IN, www.amplichip.us) detects variations in *CYP2D6* (27 alleles) and *CYP2C19* (3 alleles) (de Leon et al., 2006). Affymetrix (Santa Clara, CA) is the AmpliChip technology provider and market their own system as DMET™ Plus Premier Pack, claiming it 'features markers in all FDA-

validated genes and covers more than 90 percent of the current ADME Core markers as defined by the PharmaADME group' (www.affymetrix.com/products_services/arrays/specific/dmet.affx).

However, these products are generally based on variants of higher prevalence. Average minor allele frequencies of interrogated markers are usually 20 percent, with some below 9 percent in the DMET Plus Panel, according to Affymetrix marketing material. Including variants in the low, rare/low and rare categories, focusing on individual variability in drug response, would therefore be a main distinction of the pharmacodiagnostic kit for African populations proposed here. Technologically, this should be feasible, as the number of markers on the microarrays mentioned above is still three orders of magnitude lower than the currently most advanced whole-genome systems by Affymetrix and Illumina (San Diego, CA).

On the way from bench to bedside, 'point-of-care' technology for pharmacogenetic testing is being developed. Nanoshere's (Northbrook, IL) Verigene Warfarin Metabolism Nucleic Acid Test is an *in vitro* diagnostic for the detection and genotyping of the *CYP2C9*2* and *CYP2C9*3* alleles and an SNP of the *VKORC1* gene, from EDTA-anti-coagulated whole blood samples, as an aid in the identification of patients at risk for increased warfarin sensitivity, intended to be used on the company's Verigene System (www.nanosphere-inc.com/VerigeneWarfarinMetabolismNucleicAcidTest_4472.aspx). DNA Electronics (London, U.K.) is developing its silicon-based Genalysis™ platform for pharmacogenomics amongst other applications (www.dnae.co.uk/application.htm).

As conventional genotyping methods are based on known sequence information, they do not capture unknown variants. This is particularly unfortunate in African populations of high genetic diversity, featuring high numbers of low frequency or rare polymorphisms. Therefore, re-sequencing approaches offer the highest probability of their detection. As these methods become more affordable (see 1.3.2.1), chances of capturing individual-based SNPs and haplotypes will be enhanced. However, there will be another challenge in gauging the phenotypic impact of low frequency variants, since it would be more difficult to find individuals carrying these polymorphisms in a clinical environment.

Most novel SNPs found here, and predicted to affect enzyme expression and function, were of low prevalence, suggesting that the most frequent polymorphisms had been captured before and that any new variants found are likely to be rare and largely reflect individual variability. All variants were ranked according to prevalence (see Table 19), in order to decide on their inclusion in a pharmacodiagnostic kit for African populations, focusing specifically on drug metabolising genes and ADMET properties in Africans. While such a kit would initially be microarray-based, cheaper, easy-to-use-and-interpret versions are envisaged as 'point-of-care' technology is advancing.

CONCLUSIONS

Pharmacogenetics drives personalisation of medicine, providing highly polymorphic biomarkers. Here, the variation of drug metabolising enzyme (DME) genes was studied in the largest sample so far of representative populations from eastern, western and southern Africa.

Discovery, prevalence and functionality of DME gene variants

Novel variants were discovered in *CYP2C9*, *CYP2C19*, *CYP2D6*, *NAT2*, while the baseline prevalence of all known variants was established in eight DME genes across ten African populations. The predominance of low frequency variants, detected by re-sequencing, indicates that most common polymorphisms in DME genes are already known. Projecting on future diagnostic use, this strongly emphasises a more personalised approach to pharmacogenetics, based on individual profiling of DME gene variants. The effects of these markers on enzyme expression and function, as well as their correlation with phenotypic variation, will eventually determine their clinical significance.

Differentiation and relatedness of African populations

For the first time, frequencies of known DME alleles were used as markers for estimating population differentiation and evolutionary relatedness in a representative sample of African ethnic groups. The statistical analysis of this data revealed a clear predominance of variation *within* over variation *among* populations, emphasising again an individual-based view of African pharmacogenetic diversity.

Pharmacogenetics resources for Africa

The first pharmacogenetics database of African populations was started by systematically cataloguing DME gene variants and

pharmacogenetic studies. By ranking DME polymorphisms, a kit of African pharmacodiagnostic markers was proposed. Based on these results, pharmacogenetics resources need to be expanded in Africa, in order to facilitate the translation of pharmacogenetic research into clinical practice and to enable tailoring therapeutics to Africans.

REFERENCES

1. Abe M, Suzuki T, Deguchi T. (1993) An improved method for genotyping of N-acetyltransferase polymorphism by polymerase chain reaction. Jpn J Hum Genet 38(2):163-168.

2. Agrawal S, Khan F, Bharadwaj U. (2007) Human genetic variation studies and HLA class II loci. Int J Immunogenet 34 (4):247-252.

3. Aklillu E, Carrillo JA, Makonnen E, Hellman K, Pitarque M, Bertilsson L, Ingelman-Sundberg M. (2003) Genetic polymorphism of CYP1A2 in Ethiopians affecting induction and expression: characterization of novel haplotypes with single-nucleotide polymorphisms in intron 1. Mol Pharmacol. 64(3):659-69.

4. Aklillu E, Persson I, Bertilsson L, Johansson I, Rodrigues F, Ingelman-Sundberg M. (1996) Frequent distribution of ultrarapid metabolizers of debrisoquine in an ethiopian population carrying duplicated and multiduplicated functional CYP2D6 alleles. J Pharmacol Exp Ther 278(1):441-446.

5. Aklillu,E, Dandara,C, Bertilsson,L, Masimirembwa,C. Pharmacogenetics of Cytochrome P450s in African Populations: Clinical and Molecular Evolutionary Implications. In: Suarez-Kurtz,G, editor. Pharmacogenomics in Admixed Populations. Landes Bioscience; 2007.

6. Alfieri A, Malito E, Orru R, Fraaije MW, Mattevi A. (2008) Revealing the moonlighting role of NADP in the structure of a flavin-containing monooxygenase. Proc Natl Acad Sci 105(18):6572-6577.

7. Allabi AC, Gala JL, Desager JP, Heusterspreute M, Horsmans Y. (2003) Genetic polymorphisms of CYP2C9 and CYP2C19 in the Beninese and Belgian populations. Br J Clin Pharmacol. 56(6):653-7.

8. Allabi AC, Gala JL, Horsmans Y, Babaoglu MO, Bozkurt A, Heusterspreute M et al. (2004) Functional impact of CYP2C95, CYP2C96, CYP2C98, and CYP2C911 in vivo among black Africans. Clin Pharmacol Ther 76(2):113-118.

9. Allabi AC, Gala JL, Horsmans Y. (2005) CYP2C9, CYP2C19, ABCB1 (MDR1) genetic polymorphisms and phenytoin metabolism in a Black Beninese population. Pharmacogenet Genomics. 15(11):779-86.

10. Allabi AC, Gala JL, Horsmans Y. (2005) CYP2C9, CYP2C19, ABCB1 (MDR1) genetic polymorphisms and phenytoin metabolism in a Black Beninese population. Pharmacogenet Genomics 15(11):779-86.

11. Allison M. (2008) Is personalized medicine finally arriving? Nat Biotech 26(2):509-517.

12. Alves S, Rocha J, Amorim A, Prata MJ. (2004) Tracing the origin of the most common thiopurine methyltransferase (TPMT) variants: preliminary data from the patterns of haplotypic association with two CA repeats. Ann Hum Genet.68(Pt 4):313-23.

13. Alving AS, Carson PE, Flanagan CL, Ickes CE. (1956) Enzymatic deficiency in primaquine-sensitive erythrocytes. Science 124:484-485.

14. Al-Yahyaee S, Gaffar U, Al-Ameri MM, Qureshi M, Zadjali F, Ali BH, Bayoumi R. (2007) N-acetyltransferase polymorphism among northern Sudanese. Hum Biol. 79(4):445-52.

15. Ambrosone CB, Sweeney C, Coles BF, Thompson PA, McClure GY, Korourian S et al. (2001) Polymorphisms in glutathione S-transferases (GSTM1 and GSTT1) and survival after treatment for breast cancer. Cancer Res 61(19):7130-7135.

16. Ameyaw MM, Collie-Duguid ES, Powrie RH, Ofori-Adjei D, McLeod HL. (1999) Thiopurine methyltransferase alleles in British and Ghanaian populations. Hum Mol Genet. 8(2):367-70.

17. Anastassopoulou CG, Kostrikis LG. (2003) The impact of human allelic variation on HIV-1 disease. Curr HIV Res 1(2):185-203.

18. Antoniou AC, Spurdle AB, Sinilnikova OM, Healey S, Pooley KA, Schmutzler RK et al. (2008) Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. Am J Hum Genet 82(4):937-948.

19. Ates NA, Tamer L, Ates C, Ercan B, Elipek T, Ocal K et al. (2005) Glutathione S-transferase M1, T1, P1 genotypes and risk for development of colorectal cancer. Biochem Genet 43 (3-4):149-163.

20. Avery, O. T, MacLeod, C. M, and McCarty, M. 2-1-1944.Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. J Exp Med 79(2):137-158.

21. Baghdadi JE, Orlova M, Alter A, Ranque B, Chentoufi M, Lazrak F et al. (2006) An autosomal dominant major gene confers predisposition to pulmonary tuberculosis in adults. J Exp Med 203(7):1679-1684.

22. Balaresque PL, Ballereau SJ, Jobling MA. (2007) Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet 16(2):R134-R139.

23. Baldwin RM, Ohlsson S, Pedersen RS, Mwinyi J, Ingelman-Sundberg M, Eliasson E et al. (2008) Increased omeprazole metabolism in carriers of the CYP2C19*17 allele; a pharmacokinetic study in healthy volunteers. Br J Clin Pharmacol 65(5):767-774.

24. Bamshad M, Wooding SP. (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4(2):99-111.

25. Banerjee R, Schecter GF, Flood J, Porco TC. (2008) Extensively drug-resistant tuberculosis: new strains, new challenges. Expert Rev Anti Infect Ther 6(5):713-724.

26. Bapiro TE, Hasler JA, Ridderstrom M, Masimirembwa CM. (2002) The molecular and enzyme kinetic basis for the diminished activity of the cytochrome P450 2D6.17 (CYP2D6.17) variant. Potential implications for CYP2D6 phenotyping studies and the clinical use of CYP2D6 substrate drugs in some African populations. Biochem Pharmacol 64(9):1387-1398.

27. Barahmani N, Carpentieri S, Li XN, Wang T, Cao Y, Howe L et al. (2009) Glutathione S-transferase M1 and T1 polymorphisms may predict adverse effects after therapy in children with medulloblastoma. Neuro Oncol 11(3):292-300.

28. Barrett JC, Fry B, Maller J, Daly MJ. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-265.

29. Bartlett JM. (2005) Pharmacodiagnostic testing in breast cancer: focus on HER2 and trastuzumab therapy. Am J Pharmacogenomics 5(5):303-315.

30. Bateson, W. 4-18-1905. First mention of Genetics. Unpublished Work.

31. Bathum L, Skjelbo E, Mutabingwa TK, Madsen H, Hørder M, Brøsen K. (1999) Phenotypes and genotypes for CYP2D6 and CYP2C19 in a black Tanzanian population. Br J Clin Pharmacol. 48(3):395-401.

32. Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D et al. (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3(6):e104.

33. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E et al. (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82(5):1130-1140.

34. Bell DA, Taylor JA, Butler MA, Stephens EA, Wiest J, Brubaker LH et al. (1993) Genotype/phenotype discordance for human arylamine N-

acetyltransferase (NAT2) reveals a new slow-acetylator allele common in African-Americans. Carcinogenesis 14(8):1689-1692.

35. Belle EM, Barbujani G. (2007) Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol 133(4):1137-1146.

36. Belmont JW, Leal SM. (2005) Complex phenotypes and complex genetics: an introduction to genetic studies of complex traits. Curr Atheroscler Rep 7(3):180-187.

37. Bernal ML, Sinues B, Johansson I, McLellan RA, Wennerholm A, Dahl ML et al. (1999) Ten percent of North Spanish individuals carry duplicated or triplicated CYP2D6 genes associated with ultrarapid metabolism of debrisoquine. Pharmacogenetics 9(5):657-660.

38. Blaisdell J, Mohrenweiser H, Jackson J, Ferguson S, Coulter S, Chanas B et al. (2002) Identification and functional characterization of new potentially defective alleles of human CYP2C19. Pharmacogenetics 12(9):703-711.

39. Blum M, Demierre A, Grant DM, Heim M, Meyer UA. (1991) Molecular mechanism of slow acetylation of drugs and carcinogens in humans. Proc.natl. Acad. Sci. USA 1991;88:5237-5241.

40. Bogni A, Monshouwer M, Moscone A, Hidestrand M, Ingelman-Sundberg M, Hartung T et al. (2005) Substrate specific metabolism by polymorphic cytochrome P450 2D6 alleles. Toxicol In Vitro 19(5):621-629.

41. Boukouvala S, Fakis G. (2005) Arylamine N-acetyltransferases: what we learn from genes and genomes. Drug Metab Rev 37(3):511-564.

42. Brockmoller J, Gross D, Kerb R, Drakoulis N, Roots I. (1992) Correlation between trans-stilbene oxide-glutathione conjugation activity and the deletion mutation in the glutathione S-transferase class mu gene detected by polymerase chain reaction. Biochem Pharmacol 43(3):647-650.

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

44. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L et al. (2002) A human genome diversity cell line panel. Science 296(5566):261-262.

45. Carr A, Gross AS, Hoskins JM, Penny R, Cooper DA. (1994) Acetylation phenotype and cutaneous hypersensitivity to trimethoprim-sulphamethoxazole in HIV-infected patients. AIDS 8(3).

46. Carrington M, Dean M, Martin MP, O'Brien SJ. (1999) Genetics of HIV-1 infection: chemokine receptor CCR5 polymorphism and its consequences. Hum Mol Genet 8(10):1939-1945.

47. Casanova JL, Abel L. (2007) Human genetics of infectious diseases: a unified theory. EMBO J 26(4):915-922.

48. Cashman JR, Akerman BR, Forrest SM, Treacy EP. (2000) Population-specific polymorphisms of the human FMO3 gene: significance for detoxication. Drug Metab Dispos 28(2):169-173.

49. Cashman JR, Zhang J, Leushner J, Braun A. (2001) Population distribution of human flavin-containing monooxygenase form 3: gene polymorphisms. Drug Metab Dispos 29(12):1629-1637.

50. Cavaco I, Strömberg-Nörklit J, Kaneko A, Msellem MI, Dahoma M, Ribeiro VL, Bjorkman A, Gil JP. (2005) CYP2C8 polymorphism frequencies among malaria patients in Zanzibar. Eur J Clin Pharmacol. 61(1):15-8.

51. Cavalli-Sforza LL. (2005) The Human Genome Diversity Project: past, present and future. Nat Rev Genet 6(4):333-340.

52. Cavalli-Sforza,LL, Menozzi,P, Piazza,A. (2004) The History and Geography of Human Genes. Princeton University Press.

53. Chelule PK, Gordon M, Palanee T, Page T, Mosam A, Coovadia HM, Cassol S. (2003) MDR1 and CYP3A4 polymorphisms among African, Indian, and white populations in KwaZulu-Natal, South Africa. Clin Pharmacol Ther. 74(2):195-6.

54. Chelule PK, Pegoraro RJ, Gqaleni N, Dutton MF. (2006) The frequency of cytochrome P450 2E1 polymorphisms in Black South Africans. Dis Markers. 22(5-6):351-4.

55. Chen CJ, Yu MW, Liaw YF, Wang LW, Chiamprasert S, Matin F et al. (1996) Chronic hepatitis B carriers with null genotypes of glutathione S-transferase M1 and T1 polymorphisms who are exposed to aflatoxin are at increased risk of hepatocellular carcinoma. Am J Hum Genet 59(1):128-134.

56. Chen G, Adeyemo AA, Johnson T, Zhou J, Amoah A, Owusu S et al. (2005) A genome-wide scan for quantitative trait loci linked to obesity phenotypes among West Africans. Int J Obes (Lond) 29(3):255-259.

57. Chenevix-Trench G, Milne RL, Antoniou AC, Couch FJ, Easton DF, Goldgar DE. (2007) An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). Breast Cancer Res 9(2):104.

58. Cheng TM, Lu YE, Vendruscolo M, Lio' P, Blundell TL. (2008) Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comput Biol 4(7):e1000135.

59. Cho JY, Lim HS, Chung JY, Yu KS, Kim JR, Shin SG et al. (2004) Haplotype structure and allele frequencies of CYP2B6 in a Korean population. Drug Metab Dispos 32(12):1341-1344.

60. CIA. CIA World Factbook. 2009.

61. Cline J, Braman JC, Hogrefe HH. (1996) PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 24(18):3546-3551.

62. Collins A. (2009) Allelic association: linkage disequilibrium structure and gene mapping. Mol Biotechnol 41(1):83-89.

63. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38(11):1251-1260.

64. Cotton SC, Sharp L, Little J, Brockton N. (2000) Glutathione S-transferase polymorphisms and colorectal cancer: a HuGE review. Am J Epidemiol 151(1):7-32.

65. Crick F. (1970) Central dogma of molecular biology. Nature 227(5258):561-563.

66. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. (1961) General nature of the genetic code for proteins. Nature 192:1227-1232.

67. Crowley JJ, Sullivan PF, McLeod HL. (2009) Pharmacogenomic genome-wide association studies: lessons learned thus far. Pharmacogenomics 10(2):161-163.

68. Dandara C, Ballo R, Parker MI. (2005) CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. Cancer Lett 225(2):275-282.

69. Dandara C, Basvi PT, Bapiro TE, Sayi J, Hasler JA. (2004) Frequency of -163 C>A and 63 C>G single nucleotide polymorphism of cytochrome P450 1A2 in two African populations. Clin Chem Lab Med. 42(8):939-41.

70. Dandara C, Masimirembwa CM, Magimba A, Kaaya S, Sayi J, Sommers DK, Snyman JR, Hasler JA. (2003) Arylamine N-acetyltransferase (NAT2) genotypes in Africans: the identification of a new allele with nucleotide changes 481C>T and 590G>A. Pharmacogenetics. 13(1):55-8.

71. Dandara C, Masimirembwa CM, Magimba A, Sayi J, Kaaya S, Sommers DK, Snyman JR, Hasler JA. (2001) Genetic polymorphism of CYP2D6 and CYP2C19 in east- and southern African populations including psychiatric patients. Eur J Clin Pharmacol. 57(1):11-7.

72. Dandara C, Sayi J, Masimirembwa CM, Magimba A, Kaaya S, De Sommers K, Snyman JR, Hasler JA. (2002) Genetic polymorphism of cytochrome P450 1A1 (Cyp1A1) and glutathione transferases (M1, T1 and P1) among Africans. Clin Chem Lab Med. 40(9):952-7.

73. Danielson PB. (2002) The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr Drug Metab 3(6):561-597.

74. Danoff TM, Campbell DA, McCarthy LC, Lewis KF, Repasch MH, Saunders AM et al. (2004) A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia. Pharmacogenomics J 4(1):49-53.

75. de Graaf C, Oostenbrink C, Keizers PH, Vugt-Lussenburg BM, van Waterschoot RA, Tschirret-Guth RA et al. (2007) Molecular modeling-guided site-directed mutagenesis of cytochrome P450 2D6. Curr Drug Metab 8(1):59-77.

76. de Leon J, Susce MT, Murray-Carmichael E. (2006) The AmpliChip CYP450 genotyping test: Integrating a new clinical tool. Mol Diagn Ther 10(3):135-151.

77. de Morais SM, Wilkinson GR, Blaisdell J, Nakamura K, Meyer UA, Goldstein JA. (1994) The major genetic defect responsible for the polymorphism of S-mephenytoin metabolism in humans. J Biol Chem 269(22):15419-15422.

78. Deeni YY, Paine MJ, Ayrton AD, Clarke SE, Chenery R, Wolf CR. (2001) Expression, purification, and biochemical characterization of a human cytochrome P450 CYP2D6-NADPH cytochrome P450 reductase fusion protein. Arch Biochem Biophys 396(1):16-24.

79. DeLozier TC, Lee SC, Coulter SJ, Goh BC, Goldstein JA. (2005) Functional characterization of novel allelic variants of CYP2C9 recently discovered in southeast Asians. J Pharmacol Exp Ther 315(3):1085-1090.

80. Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schafer H. (2008) Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 16(10):1164-1172.

81. Dickinson GL, Lennard MS, Tucker GT, Rostami-Hodjegan A. (2007) The use of mechanistic DM-PK-PD modelling to assess the power of pharmacogenetic studies -CYP2C9 and warfarin as an example. Br J Clin Pharmacol 64(1):14-26.

82. Dickmann LJ, Rettie AE, Kneller MB, Kim RB, Wood AJ, Stein CM et al. (2001) Identification and functional characterization of a new CYP2C9 variant (CYP2C9*5) expressed among African Americans. Mol Pharmacol 60(2):382-387.

83. Donald PR, Parkin DP, Seifart HI, Schaaf HS, van Helden PD, Werely CJ et al. (2007) The influence of dose and N-acetyltransferase-2 (NAT2) genotype and phenotype on the pharmacokinetics and pharmacodynamics of isoniazid. Eur J Clin Pharmacol 63(7):633-639.

84. Droll K, Bruce-Mensah K, Otton SV, Gaedigk A, Sellers EM, Tyndale RF. (1998) Comparison of three CYP2D6 probe substrates and genotype in Ghanaians, Chinese and Caucasians. Pharmacogenetics. 8(4):325-33.

85. Ebid AH, Ahmed MM, Mohammed SA. (2007) Therapeutic drug monitoring and clinical outcomes in epileptic Egyptian patients: a gene polymorphism perspective study. Ther Drug Monit. 29(3):305-12.

86. Ekhart C, Rodenhuis S, Smits PH, Beijnen JH, Huitema AD. (2009) An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug treatment. Cancer Treat Rev 35(1):18-31.

87. Epstein, D. and Thenabadu, R. 2-25-2009. Drug resistance could set back malaria control success. WHO Media Centre

88. Evans DA, Manley KA, Mckusick VA. (1960) Genetic control of isoniazid metabolism in man. Br Med J 2(5197):485-491.

89. Evans WE, Relling MV. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286(5439):487-491.

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

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

92. FDA, 8-16-2007. FDA Approves Updated Warfarin (Coumadin) Prescribing Information: New Genetic Information May Help Providers Improve Initial Dosing Estimates of the Anticoagulant for Individual Patients. U.S.Food and Drug Administration.

93. FDA. 5-1-2009. Plavix (clopidogrel bisulfate) 75 mg tablets: Safety Labeling Changes Approved By FDA Center for Drug Evaluation and Research (CDER). U.S.Food and Drug Administration.

94. FDA. 7-8-2008. FDA Approves New Genetic Test for Patients with Breast Cancer. U.S.Food and Drug Administration.

95. FDA. 8-22-2005. FDA Clears Genetic Test That Advances Personalized Medicine Test Helps Determine Safety of Drug Therapy. U.S.Food and Drug Administration.

96. Feero WG, Guttmacher AE, Collins FS. (2008) The genome gets personal--almost. JAMA 299(11):1351-1352.

97. Ferrara J. (2007) Personalized medicine: challenging pharmaceutical and diagnostic company business models. Mcgill J Med 10(1):59-61.

98. Ferreira PE, Veiga MI, Cavaco I, Martins JP, Andersson B, Mushin S, Ali AS, Bhattarai A, Ribeiro V, Björkman A, Gil JP. (2008) Polymorphism of antimalaria drug metabolizing, nuclear receptor, and drug transport genes among malaria patients in Zanzibar, East Africa. Ther Drug Monit. 30(1):10-5.

99. Feuk L, Carson AR, Scherer SW. (2006) Structural variation in the human genome. Nat Rev Genet 7(2):85-97.

100. Fisher RA. (1922) On the interpretation of 2x2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85(1):87-94.

101. Fitness J, Floyd S, Warndorff DK, Sichali L, Malema S, Crampin AC et al. (2004) Large-scale candidate gene study of tuberculosis susceptibility in the Karonga district of northern Malawi. Am J Trop Med Hyg 71(3):341-349.

102. Fox, A. F. 1932.The relationship between chemical constitution and taste. Genetics 18, 115-120.

103. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM et al. (2006) Copy number variation: new insights in genome diversity. Genome Res 16(8):949-961.

104. Frost and Sullivan. 6-4-2009. Kenyan Pharmaceutical Industry Can Expect Significant Growth Despite Global Conditions.

105. Fung KL, Gottesman MM. (2009) A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. Biochim Biophys Acta 1794(5):860-871.

106. Furuta T, Shirai N, Kodaira M, Sugimoto M, Nogaki A, Kuriyama S et al. (2007) Pharmacogenomics-based tailored versus standard therapeutic regimen for eradication of H. pylori. Clin Pharmacol Ther 81(4):521-528.

107. Gaedigk A, Blum M, Gaedigk R, Eichelbaum M, Meyer UA. (1991) Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. Am J Hum Genet 48(5):943-950.

108. Gaedigk A, Gotschall RR, Forbes NS, Simon SD, Kearns GL, Leeder JS. (1999) Optimization of cytochrome P4502D6 (CYP2D6) phenotype assignment using a genotyping algorithm based on allele frequency data. Pharmacogenetics 9(6):669-682.

109. Gaedigk A, Simon SD, Pearce RE, Bradford LD, Kennedy MJ, Leeder JS. (2008) The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. Clin Pharmacol Ther 83(2):234-242.

110. Garret RH, Grisham,CM. Biochemistry. Forth Worth: Saunders College Publications; 1995.

111. Garrigan D, Kingan SB, Pilkington MM, Wilder JA, Cox MP, Soodyall H et al. (2007) Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics 177(4):2195-2207.

112. Garrod,AE. The inborn errors of metabolism. London: Oxford University Press; 1909.

113. Germano S, O'Driscoll L. (2009) Breast cancer: understanding sensitivity and resistance to chemotherapy and targeted therapies to aid in personalised medicine. Curr Cancer Drug Targets 9(3):398-418.

114. Giacomini KM, Brett CM, Altman RB, Benowitz NL, Dolan ME, Flockhart DA et al. (2007) The pharmacogenetics research network: from SNP discovery to clinical drug response. Clin Pharmacol Ther 81(3):328-345.

115. Gisbert JP, Nino P, Rodrigo L, Cara C, Guijarro LG. (2006) Thiopurine methyltransferase (TPMT) activity and adverse effects of azathioprine in inflammatory bowel disease: long-term follow-up study of 394 patients. Am J Gastroenterol 101(12):2769-2776.

116. Goetz MP, Knox SK, Suman VJ, Rae JM, Safgren SL, Ames MM et al. (2007) The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. Breast Cancer Res Treat 101(1):113-121.

117. Goldstein DB. (2005) The genetics of human drug response. Philos Trans R Soc Lond B Biol Sci 360(1460):1571-1572.

118. Goldstein JA, Ishizaki T, Chiba K, de Morais SM, Bell D, Krahn PM et al. (1997) Frequencies of the defective CYP2C19 alleles responsible for the mephenytoin poor metabolizer phenotype in various Oriental, Caucasian, Saudi Arabian and American black populations. Pharmacogenetics 7(1):59-64.

119. Goldstein JA. (2001) Clinical relevance of genetic polymorphisms in the human CYP2C subfamily. Br J Clin Pharmacol 52(4):349-355.

120. Gondos A, Chokunonga E, Brenner H, Parkin DM, Sankila R, Borok MZ et al. (2004) Cancer survival in a southern African urban population. Int J Cancer 112(5):860-864.

121. Gong L, Owen RP, Gor W, Altman RB, Klein TE. (2008) PharmGKB: an integrated resource of pharmacogenomic data and knowledge. Curr Protoc Bioinformatics Chapter 14:Unit14.

122. Gordon,RG. Ethnologue: Languages of the World.16th edition. Dallas, Texas: SIL International; 2005.

123. Gotoh O. (1992) Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. J Biol Chem 267(1):83-90.

124. Gray IC, Nobile C, Muresu R, Ford S, Spurr NK. (1995) A 2.4-megabase physical map spanning the CYP2C gene cluster on chromosome 10q24. Genomics 28(2):328-332.

125. Griese EU, Asante-Poku S, Ofori-Adjei D, Mikus G, Eichelbaum M. (1999) Analysis of the CYP2D6 gene mutations and their consequences for enzyme function in a West African population. Pharmacogenetics. 9(6):715-23.

126. Grossman I. (2007) Routine pharmacogenetic testing in clinical practice: dream or reality? Pharmacogenomics 8(10):1449-1459.

127. Guengerich FP. (2003) Cytochromes P450, drugs, and diseases. Mol Interv 3(4):194-204.

128. Guessous I, Gwinn M, Khoury MJ. (2009) Genome-wide association studies in pharmacogenomics: untapped potential for translation. Genome Med 1(4):46.

129. Gurwitz D, McLeod HL. (2009) Genome-wide association studies: powerful tools for improving drug safety and efficacy. Pharmacogenomics 10(2):157-159.

130. Guttmacher AE, Collins FS. (2002) Genomic medicine--a primer. N Engl J Med 347(19):1512-1520.

131. Haas DW, Bartlett JA, Andersen JW, Sanne I, Wilkinson GR, Hinkle J, Rousseau F, Ingram CD, Shaw A, Lederman MM, Kim RB; Adult AIDS Clinical Trials Group. (2006) Pharmacogenetics of nevirapine-associated hepatotoxicity: an Adult AIDS Clinical Trials Group collaboration. Clin Infect Dis. 2006 43(6):783-6.

132. Haas DW, Ribaudo HJ, Kim RB, Tierney C, Wilkinson GR, Gulick RM et al. (2004) Pharmacogenetics of efavirenz and central nervous system side effects: an Adult AIDS Clinical Trials Group study. AIDS 18(18):2391-2400.

133. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 29(5):648-658.

134. Hamdy SI, Hiratsuka M, Narahara K, El-Enany M, Moursi N, Ahmed MS, Mizugaki M. (2002) Allele and genotype frequencies of polymorphic cytochromes P450 (CYP2C9, CYP2C19, CYP2E1) and dihydropyrimidine dehydrogenase (DPYD) in the Egyptian population. Br J Clin Pharmacol. 53(6):596-603.

135. Hamdy SI, Hiratsuka M, Narahara K, Endo N, El-Enany M, Moursi N, Ahmed MS, Mizugaki M. (2003) Genotype and allele frequencies of TPMT, NAT2, GST, SULT1A1 and MDR-1 in the Egyptian population. Br J Clin Pharmacol. 55(6):560-9.

136. Hamdy SI, Hiratsuka M, Narahara K, Endo N, El-Enany M, Moursi N, Ahmed MS, Mizugaki M. (2003) Genotyping of four genetic polymorphisms in the CYP1A2 gene in the Egyptian population. Br J Clin Pharmacol. 55(3):321-4.

137. Hammer SM, Eron JJ, Jr., Reiss P, Schooley RT, Thompson MA, Walmsley S et al. (2008) Antiretroviral treatment of adult HIV infection: 2008 recommendations of the International AIDS Society-USA panel. JAMA 300(5):555-570.

138. Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R. (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 55(1):175-189.

139. Handley LJ, Manica A, Goudet J, Balloux F. (2007) Going the distance: human population genetics in a clinal world. Trends Genet 23(9):432-439.

140. Hanioka N, Kimura S, Meyer UA, Gonzalez FJ. (1990) The human CYP2D locus associated with a common genetic defect in drug oxidation: a G1934----A base change in intron 3 of a mutant CYP2D6 allele results in an aberrant 3' splice recognition site. Am J Hum Genet 47(6):994-1001.

141. Hao D, Sun J, Furnes B, Schlenk D, Li M, Yang S et al. (2007) Allele and genotype frequencies of polymorphic FMO3 gene in two genetically distinct populations. Cell Biochem Funct 25(4):443-453.

142. Hardy BJ, Seguin B, Ramesar R, Singer PA, Daar AS. (2008) South Africa: from species cradle to genomic applications. Nat Rev Genet 9(1):S19-S23.

143. Hardy GH. (1908) Mendelian proportions in a mixed Population. Science 28(706):49-50.

144. Hardy J, Singleton A. (2009) Genomewide association studies and human disease. N Engl J Med 360(17):1759-1768.

145. Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST. (1998) Genetic traces of ancient demography. Proc Natl Acad Sci USA 95(4):1961-1967.

146. Hartl,DA, Clark,AG. Principles of population Genetics. Sunderland, MA: Sinauer Associates, 1997.

147. Heckmann JM, Lambson EM, Little F, Owen EP. (2005) Thiopurine methyltransferase (TPMT) heterozygosity and enzyme activity as predictive tests for the development of azathioprine-related adverse events. J Neurol Sci. 231(1-2):71-80.

148. Hein DW, Boukouvala S, Grant DM, Minchin RF, Sim E. (2008) Changes in consensus arylamine N-acetyltransferase gene nomenclature. Pharmacogenet Genomics 18(4):367-368.

149. Hein R, Beckmann L, Chang-Claude J. (2008) Sample size requirements for indirect association studies of gene-environment interactions (G x E). Genet Epidemiol 32(3):235-245.

150. Herrlin K, Massele AY, Jande M, Alm C, Tybring G, Abdi YA et al. (1998) Bantu Tanzanians have a decreased capacity to metabolize omeprazole and mephenytoin in relation to their CYP2C19 genotype. Clin Pharmacol Ther 64(4):391-401.

151. Hert DG, Fredlake CP, Barron AE. (2008) Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29(23):4618-4626.

152. Higashi MK, Veenstra DL, Kondo LM, Wittkowsky AK, Srinouanprachanh SL, Farin FM et al. (2002) Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. JAMA 287(3):1690-1698.

153. Hiratsuka M, Takekuma Y, Endo N, Narahara K, Hamdy SI, Kishikawa Y et al. (2002) Allele and genotype frequencies of CYP2B6 and CYP3A5 in the Japanese population. Eur J Clin Pharmacol 58(6):417-421.

154. Hofmann MH, Blievernicht JK, Klein K, Saussele T, Schaeffeler E, Schwab M et al. (2008) Aberrant splicing caused by single nucleotide polymorphism c.516G>T [Q172H], a marker of CYP2B6*6, is responsible for decreased expression and activity of CYP2B6 in liver. J Pharmacol Exp Ther 325(1):284-292.

155. Hood LE, Gallas DJ. (2008) P4 Medicine: Personalized, Predictive, Preventive, Participatory: A Change of View that Changes Everything. http://www.cra.org/ccc/docs/init/P4_Medicine.pdf accessed 6 March 2009.

156. Hung CC, Lin CJ, Chen CC, Chang CJ, Liou HH. (2004) Dosage recommendation of phenytoin for patients with epilepsy with different CYP2C9/CYP2C19 polymorphisms. Ther Drug Monit 26(5):534-40.

157. Ibeanu GC, Goldstein JA, Meyer U, Benhamou S, Bouchardy C, Dayer P et al. (1998) Identification of new human CYP2C19 alleles (CYP2C19*6 and CYP2C19*2B) in a Caucasian poor metabolizer of mephenytoin. J Pharmacol Exp Ther 286(3):1490-1495.

158. Ingelman-Sundberg M, Oscarson M, Daly AK, Garte S, and Nebert DW, 2001.Human cytochrome P-450 (CYP) genes: a web page for the nomenclature of alleles. Cancer Epidemiol Biomarkers Prev. 10(12):1307-1308.

159. Ingelman-Sundberg M, Rodriguez-Antona C. (2005) Pharmacogenetics of drug-metabolizing enzymes: implications for a safer and more effective drug therapy. Philos Trans R Soc Lond B Biol Sci 360(1460):1563-1570.

160. Ingelman-Sundberg M. (2005) Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J 5(1):6-13.

161. Isin EM, Guengerich FP. (2008) Substrate binding to cytochromes P450. Anal Bioanal Chem 392(6):1019-1030.

162. Ivanovic J, Nicastri E, Ascenzi P, Bellagamba R, De Marinis E, Notari S et al. (2008) Therapeutic drug monitoring in the management of HIV-infected patients. Curr Med Chem 15(19):1925-1939.

163. Jackson, F. 6-3-2006. New Advances in Molecular Anthropological Genetics. Conference Proceeding.

164. Johnson JA, Boerwinkle E, Zineh I, Chapman AB, Bailey K, Cooper-DeHoff RM et al. (2009) Pharmacogenomics of antihypertensive drugs: rationale and design of the Pharmacogenomic Evaluation of Antihypertensive Responses (PEAR) study. Am Heart J 157(3):442-449.

165. Jorde, L. B and Olson, S. 2008.Race, Genetics, and Healthcare. National Coalition for Health Professional Education in Genetics.Electronic Citation

166. Joshua AM, Boutros PC. (2008) Web-based resources for clinical bioinformatics. Methods Mol Med 141:309-329.

167. Kahn J. (2008) Exploiting race in drug development: BiDil's interim model of pharmacogenomics. Soc Stud Sci 38(5):737-758.

168. Kaiser J. (2003) Genomic medicine. African-American population biobank proposed. Science 300(5625):1485.

169. Kancko A, Bergqvist Y, Taleo G, Kobayakawa T, Ishizaki T, Bj+¦rkman A. (1999) Proguanil disposition and toxicity in malaria patients from Vanuatu with high frequencies of CYP2C19 mutations. Pharmacogenetics and Genomics 9(3).

170. Karchin R. (2009) Next generation tools for the annotation of human SNPs. Brief Bioinform 10(1):35-52.

171. Kealey C, Chen Z, Christie J, Thorn CF, Whitehead AS, Price M et al. (2007) Warfarin and cytochrome P450 2C9 genotype: possible ethnic variation in warfarin sensitivity. Pharmacogenomics 8(3):217-225.

172. Kidd RS, Curry TB, Gallagher S, Edeki T, Blaisdell J, Goldstein JA. (2001) Identification of a null allele of CYP2C9 in an African-American exhibiting toxicity to phenytoin. Pharmacogenetics 11(9):803-808.

173. Kim HR, Hwang SS, Kim HJ, Lee SM, Yoo CG, Kim YW et al. (2007) Impact of extensive drug resistance on treatment outcomes in non-HIV-infected patients with multidrug-resistant tuberculosis. Clin Infect Dis 45(10):1290-1295.

174. Kimchi-Sarfaty C, Marple AH, Shinar S, Kimchi AM, Scavo D, Roma MI et al. (2007) Ethnicity-related polymorphisms and haplotypes in the human ABCB1 gene. Pharmacogenomics 8(1):29-39.

175. Kimura S, Umeno M, Skoda RC, Meyer UA, Gonzalez FJ. (1989) The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. Am J Hum Genet 45(6):889-904.

176. Kindmark A, Jawaid A, Harbron CG, Barratt BJ, Bengtsson OF, Andersson TB et al. (2007) Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis. Pharmacogenomics J 8(3):186-195.

177. Kirchheiner J, Rodriguez-Antona C. (2009) Cytochrome P450 2D6 genotyping: potential role in improving treatment outcomes in psychiatric disorders. CNS Drugs 23(3):181-191.

178. Kirchheiner J. (2008) CYP2D6 phenotype prediction from genotype: which system is the best? Clin Pharmacol Ther 83(2):225-227.

179. Kitada M. (2003) Genetic polymorphism of cytochrome P450 enzymes in Asian populations: focus on CYP2D6. Int J Clin Pharmacol Res 23(1):31-35.

180. Klein K, Lang T, Saussele T, Barbosa-Sicard E, Schunck WH, Eichelbaum M, Schwab M, Zanger UM. (2005) Genetic variability of CYP2B6 in populations of African and Asian origin: allele frequencies, novel functional variants, and possible implications for anti-HIV therapy with efavirenz. Pharmacogenet Genomics. 15(12):861-73.

181. Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT et al. (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 360(8):753-764.

182. Koukouritaki SB, Hines RN. (2005) Flavin-containing monooxygenase genetic polymorphism: impact on chemical metabolism and drug development. Pharmacogenomics 6:807-822.

183. Koukouritaki SB, Poch MT, Henderson MC, Siddens LK, Krueger SK, VanDyke JE et al. (2007) Identification and functional analysis of common human flavin-containing monooxygenase 3 genetic variants. J Pharmacol Exp Ther 320(1):266-273.

184. Kraft P, Hunter DJ. (2009) Genetic risk prediction--are we there yet? N Engl J Med 360(17):1701-1703.

185. Krueger SK, Williams DE. (2005) Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. Pharmacol Ther 106(3):357-387.

186. Kuznetsov IB, McDuffie M, Moslehi R. (2009) A web server for inferring the human N-acetyltransferase-2 (NAT2) enzymatic phenotype from NAT2 genotype. Bioinformatics 25(9):1185-1186.

187. Laberge,A-M, Burke,W. (2008) Personalized Medicine and Genomics. In: Crowley,M, editor. From Birth to Death and Bench to Clinic: The Hastings Center Bioethics Briefing Book for Journalists, Policymakers, and Campaigns. New York: Garrison; pp. 133-136.

188. Lamba V, Lamba J, Yasuda K, Strom S, Davila J, Hancock ML et al. (2003) Hepatic CYP2B6 expression: gender and ethnic differences and relationship to CYP2B6 genotype and CAR (constitutive androstane receptor) expression. J Pharmacol Exp Ther 307(3):906-922.

189. Lang T, Klein K, Richter T, Zibat A, Kerb R, Eichelbaum M et al. (2004) Multiple novel nonsynonymous CYP2B6 gene polymorphisms in Caucasians: demonstration of phenotypic null alleles. J Pharmacol Exp Ther 311(1):34-43.

190. Lattard V, Zhang J, Tran Q, Furnes B, Schlenk D, Cashman JR. (2003) Two new polymorphisms of the FMO3 gene in Caucasian and African-

American populations: comparative genetic and functional studies. Drug Metab Dispos 31(7):854-860.

191. Laurent C, Bourgeois A, Mpoudi-Ngole E, Ciaffi L, Kouanfack C, Mougnutou R et al. (2008) Tolerability and effectiveness of first-line regimens combining nevirapine and lamivudine plus zidovudine or stavudine in Cameroon. AIDS Res Hum Retroviruses 24(3):393-399.

192. Ledesma MC, Agundez JA. (2005) Identification of subtypes of CYP2D gene rearrangements among carriers of CYP2D6 gene deletion and duplication. Clin Chem 51(6):939-943.

193. Lee KM, Park SK, Kim SU, Doll MA, Yoo KY, Ahn SH et al. (2003) N-acetyltransferase (NAT1, NAT2) and glutathione S-transferase (GSTM1, GSTT1) polymorphisms in breast cancer. Cancer Lett 196(2):179-186.

194. Lees P, Cunningham FM, Elliott J. (2004) Principles of pharmacodynamics and their applications in veterinary pharmacology. J Vet Pharmacol Ther 27(6):397-414.

195. Lehmann H, Ryan E. (1956) The familial incidence of low pseudocholinesterase level. Lancet 271(6934):124.

196. Lesko LJ. (2007) Personalized medicine: elusive dream or imminent reality? Clin Pharmacol Ther 81(6):807-816.

197. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100-1104.

198. Lin J, Kamat A, Gu J, Chen M, Dinney CP, Forman MR et al. (2009) Dietary Intake of Vegetables and Fruits and the Modification Effects of GSTM1 and NAT2 Genotypes on Bladder Cancer Risk. Cancer Epidemiol Biomarkers Prev.

199. Lindh JD, Holm L, Andersson ML, Rane A. (2009) Influence of CYP2C9 genotype on warfarin dose requirements--a systematic review and meta-analysis. Eur J Clin Pharmacol 65(4):365-375.

200. London WT, Evans AA, Buetow K, Litwin S, McGlynn K, Zhou T et al. (1995) Molecular and genetic epidemiology of hepatocellular carcinoma: studies in China and Senegal. Princess Takamatsu Symp 25:51-60.

201. Lundberg KS, Shoemaker DD, Adams MW, Short JM, Sorge JA, Mathur EJ. (1991) High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. Gene 108(1):1-6.

202. Mahgoub A, Idle JR, Dring LG, Lancaster R, Smith RL. (1977) Polymorphic hydroxylation of debrisoquine in man. Lancet ii:584-586.

203. Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C et al. (2002) Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet 359(9308):727-732.

204. Mamiya K, Ieiri I, Miyahara S, Imai J, Furuumi H, Fukumaki Y et al. (1998) Association of polymorphisms in the cytochrome P450 (CYP) 2C19 and 2C18 genes in Japanese epileptic patients. Pharmacogenetics 8(1):87-90.

205. Manfredi S, Calvi D, del Fiandra M, Botto N, Biagini A, Andreassi MG. (2009) Glutathione S-transferase T1- and M1-null genotypes and coronary artery disease risk in patients with Type 2 diabetes mellitus. Pharmacogenomics 10(1):29-34.

206. Manolio TA, Brooks LD, Collins FS. (2008a) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118(5):1590-1605.

207. Mantel N. (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27(2):209-220.

208. Mao, M., Matimba, A., Scordo, M. G., Günesa, A., Zengile, H., Yasui-Furukori, N. et al. 2008.Common FMO3 polymorphisms in 13 ethnic populations from Europe, East Asia and sub-Saharan Africa: frequency and linkage analysis. Vienna, Austria. Conference Proceeding.

209. Mardis ER. (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387-402.

210. Marioni JC, White M, Tavare S, Lynch AG. (2008) Hidden copy number variation in the HapMap population. Proc Natl Acad Sci USA 105(29):10067-10072.

211. Martinez FD, Graves PE, Baldini M, Solomon S, Erickson R. (1997) Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. J Clin Invest 100(12):3184-3188.

212. Masimirembwa C, Bertilsson L, Johansson I, Hasler JA, Ingelman-Sundberg M. (1995) Phenotyping and genotyping of S-mephenytoin hydroxylase (cytochrome P450 2C19) in a Shona population of Zimbabwe. Clin Pharmacol Ther 57:656-661.

213. Masimirembwa C, Persson I, Bertilsson L, Hasler J, Ingelman-Sundberg M. (1996) A novel mutant variant of the CYP2D6 gene (CYP2D6*17) common in a black African population: association with diminished debrisoquine hydroxylase activity. Br J Clin Pharmacol 42(6):713-719.

214. Masimirembwa CM, Dandara C, Sommers DK, Snyman JR, Hasler JA. (1998) Genetic polymorphism of cytochrome P4501A1, microsomal epoxide hydrolase, and glutathione S-transferases M1 and T1 in Zimbabweans and Venda of southern Africa. Pharmacogenetics. 8(1):83-5.

215. Matimba A, Del Favero J, Van Broeckhoven C, Masimirembwa C. (2009) Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects. Hum Genomics 3(2):169-190.

216. Matimba A, Oluka MN, Ebeshi BU, Sayi J, Bolaji OO, Guantai AN et al. (2008) Establishment of a biobank and pharmacogenetics database of African populations. Eur J Hum Genet.

217. McCarroll SA. (2008) Extending genome-wide association studies to copy-number variation. Hum Mol Genet 17(R2):R135-R142.

218. McGrath M, Michaud D, De V, I. (2006) Polymorphisms in GSTT1, GSTM1, NAT1 and NAT2 genes and bladder cancer risk in men and women. BMC Cancer 6(6):239.

219. McKusick,VA. Mendelian Inheritance in Man.12th edition. Baltimore: Johns Hopkins Press; 1998.

220. McLary, D. 5-17-2009.Cancer Rates Rising in Africa. Voice of America Newspaper.

221. McLellan RA, Oscarson M, Alexandrie AK, Seidegard J, EVANS DA, Rannug A et al. (1997) Characterization of a human glutathione S-transferase mu cluster containing a duplicated GSTM1 gene that causes ultrarapid enzyme activity. Mol Pharmacol 52(6):958-965.

222. McLeod HL, Pritchard SC, Githang'a J, Indalo A, Ameyaw MM, Powrie RH, Booth L, Collie-Duguid ES. Ethnic differences in thiopurine methyltransferase pharmacogenetics: evidence for allele specificity in Caucasian and Kenyan individuals. (1999) Pharmacogenetics. 9(6):773-6.

223. Mehlotra RK, Bockarie MJ, Zimmerman PA. (2007) CYP2B6 983T>C polymorphism is prevalent in West Africa but absent in Papua New Guinea: implications for HIV/AIDS treatment. Br J Clin Pharmacol 64(3):391-395.

224. Mehlotra RK, Ziats MN, Bockarie MJ, Zimmerman PA. (2006) Prevalence of CYP2B6 alleles in malaria-endemic populations of West Africa and Papua New Guinea. Eur J Clin Pharmacol. 62(4):267-75.

225. Mehta U, Durrheim DN, Blockman M, Kredo T, Gounden R, Barnes KI. (2007) Adverse drug reactions in adult medical inpatients in a South African hospital serving a community with a high HIV/AIDS prevalence: prospective observational study. Br J Clin Pharmacolol 65(3):396-406

226. Mestres J. (2005) Structure conservation in cytochromes P450. Proteins 58(3):596-609.

227. Meyer UA. (2004) Pharmacogenetics - five decades of therapeutic lessons from genetic diversity. Nat Rev Genet 5(9):669-676.

228. Minami H, Sai K, Saeki M, Saito Y, Ozawa S, Suzuki K et al. (2007) Irinotecan pharmacokinetics/pharmacodynamics and UGT1A genetic polymorphisms in Japanese: roles of UGT1A1*6 and *28. Pharmacogenet Genomics 17(7):497-504.

229. Miners JO, Birkett DJ. (1998) Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. Br J Clin Pharmacol 45(6):525-538.

230. Mitchell RS, Bell JC. (1957) Clinical implications of isoniazid, PAS and streptomycin blood levels in tuberculosis. Trans Am Clin Climatol Assoc 69:98-102-105.

231. Mo Z, Gao Y, Cao Y, Gao F, Jian L. (2009) An updating meta-analysis of the GSTM1, GSTT1, and GSTP1 polymorphisms and prostate cancer: a HuGE review. Prostate 69(6):662-688.

232. Motulsky AG. (2006) Genetics of complex diseases. J Zhejiang Univ Sci B 7(2):167-168.

233. Ng PC, Henikoff S. (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812-3814.

234. Nolan D, Phillips E, Mallal S. (2006) Efavirenz and CYP2B6 polymorphism: Implications for drug toxicity and resistance. Clin Infect Dis 42(3):408-410.

235. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A et al. (2008) Genes mirror geography within Europe. Nature 456(7218):98-101.

236. Novembre J, Stephens M. (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40(5):646-649.

237. Nyakutira C, Roshammar D, Chigutsa E, Chonzi P, Ashton M, Nhachi C et al. (2008) High prevalence of the CYP2B6 516G-->T(*6) variant and effect on the population pharmacokinetics of efavirenz in HIV/AIDS outpatients in Zimbabwe. Eur J Clin Pharmacol 64(4):357-365.

238. Ohno M, Yamamoto A, Ono A, Miura G, Funamoto M, Takemoto Y et al. (2009) Influence of clinical and genetic factors on warfarin dose requirements among Japanese patients. Eur J Clin Pharmacol.

239. Oliveira E, Quental S, Alves S, Amorim A, Prata MJ. (2007) Do the distribution patterns of polymorphisms at the thiopurine S-methyltransferase locus in sub-Saharan populations need revision? Hints from Cabinda and Mozambique. Eur J Clin Pharmacol. 63(7):703-6.

240. O'Shaughnessy KM. (2009) Dissecting complex traits: recent advances in hypertension genomics. Genome Med 1(4):43.

241. Pande M, Amos CI, Osterwisch DR, Chen J, Lynch PM, Broaddus R et al. (2008) Genetic variation in genes for the xenobiotic-metabolizing enzymes CYP1A1, EPHX1, GSTM1, GSTT1, and GSTP1 and susceptibility to colorectal cancer in Lynch syndrome. Cancer Epidemiol Biomarkers Prev 17(9):2393-2401.

242. Panserat S, Sica L, Gérard N, Mathieu H, Jacqz-Aigrain E, Krishnamoorthy R. (1999) CYP2D6 polymorphism in a Gabonese population: contribution of the CYP2D6*2 and CYP2D6*17 alleles to the high prevalence of the intermediate metabolic phenotype. Br J Clin Pharmacol. 47(1):121-4.

243. Parikh S, Ouedraogo JB, Goldstein JA, Rosenthal PJ, Kroetz DL. (2007) Amodiaquine metabolism is impaired by common polymorphisms in CYP2C8: implications for malaria treatment in Africa. Clin Pharmacol Ther. 82(2):197-203.

244. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R et al. (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63(6):1839-1851.

245. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW et al. (2007) PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 3(9):1672-1686.

246. Passetti F, Ferreira CG, Costa FF. (2009) The impact of microRNAs and alternative splicing in pharmacogenomics. Pharmacogenomics J 9(1):1-13.

247. Patin E, Harmant C, Kidd KK, Kidd J, Froment A, Mehdi SQ, Sica L, Heyer E, Quintana-Murci L. (2006) Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene. Hum Mutat. 27(7):720.

248. Pearson K. (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. Philosophical Magazine 50:157-175.

249. Pearson WR, Vorachek WR, Xu SJ, Berger R, Hart I, Vannais D et al. (1993) Identification of class-mu glutathione transferase genes GSTM1-GSTM5 on human chromosome 1p13. Am J Hum Genet 53(1):220-233.

250. Peltonen L, McKusick VA. (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. Science 291(5507):1224-1229.

251. Pemble S, Schroeder KR, Spencer SR, Meyer DJ, Hallier E, Bolt HM et al. (1994) Human glutathione S-transferase theta (GSTT1): cDNA cloning and the characterization of a genetic polymorphism. Biochem J 300 (Pt 1):271-276.

252. Penzak SR, Kabuye G, Mugyenyi P, Mbamanya F, Natarajan V, Alfaro RM, Kityo C, Formentini E, Masur H. (2007) Cytochrome P450 2B6 (CYP2B6) G516T influences nevirapine plasma concentrations in HIV-infected patients in Uganda. HIV Med. 8(2):86-91.

253. Perera MA, Innocenti F, Ratain MJ. (2008) Pharmacogenetic testing for uridine diphosphate glucuronosyltransferase 1A1 polymorphisms: are we there yet? Pharmacotherapy 28(6):755-768.

254. Persson I, Aklillu E, Rodrigues F, Bertilsson L, Ingelman-Sundberg M. (1996) S-mephenytoin hydroxylation phenotype and CYP2C19 genotype among Ethiopians. Pharmacogenetics. 6(6):521-6.

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

256. Rae JM, Sikora MJ, Henry NL, Li L, Kim S, Oesterreich S et al. (2009) Cytochrome P450 2D6 activity predicts discontinuation of tamoxifen therapy in breast cancer patients. Pharmacogenomics J.

257. Rahemtulla T, Bhopal R. (2005) Pharmacogenetics and ethnically targeted therapies. BMJ 330(7499):1036-1037.

258. Rajewsky N. (2006) microRNA target predictions in animals. Nat Genet 38 Suppl:S8-13.
259. Ramensky V, Bork P, Sunyaev S. (2002a) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894-3900.

260. Raymond M, Rousset F. (1995) GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. J Heredity 86(3):248-249.

261. Raymond M, Rousset R. (1995) An exact test for population differentiation. Evolution 49(6):1280-1283.

262. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al. (2006) Global variation in copy number in the human genome. Nature 444(7118):444-454.

263. Reich D, Price AL, Patterson N. (2008) Principal component analysis of genetic data. Nat Genet 40(5):491-492.

264. Relethford JH. (2004) Global patterns of isolation by distance based on genetic and morphological data. Hum Biol 76(4):499-513.

265. Reynolds J, Weir BS, Cockerham CC. (1983) Estimating for coancestry coefficient: basis for a short-term genetic distance. Genetics 105(767):779.

266. Rogan PK, Svojanovsky S, Leeder JS. (2003) Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics 13(4):207-218.

267. Roses AD, Saunders AM, Huang Y, Strum J, Weisgraber KH, Mahley RW. (2007) Complex disease-associated pharmacogenetics: drug efficacy, drug safety, and confirmation of a pathogenetic hypothesis (Alzheimer's disease). Pharmacogenomics J 7(1):10-28.

268. Roses AD. (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. Nat Rev Genet 5(9):645-656.

269. Rotger M, Colombo S, Furrer H, Bleiber G, Buclin T, Lee BL et al. (2005) Influence of CYP2B6 polymorphism on plasma and intracellular concentrations and toxicity of efavirenz and nevirapine in HIV-infected patients. Pharmacogenet Genomics 15(1):1-5.

270. Rotger M, Saumoy M, Zhang K, Flepp M, Sahli R, Decosterd L et al. (2007) Partial deletion of CYP2B6 owing to unequal crossover with CYP2B7. Pharmacogenet Genomics 17(10):885-890.

271. Rousset F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145(4):1219-1228.

272. Rowland P, Blaney FE, Smyth MG, Jones JJ, Leydon VR, Oxbrow AK et al. (2006) Crystal structure of human cytochrome P450 2D6. J Biol Chem 281(11):7614-7622.

273. Roy PD, Majumder M, Roy B. (2008) Pharmacogenomics of anti-TB drugs-related hepatotoxicity. Pharmacogenomics 9(3):311-321.

274. Rozen S, Skaletsky H. (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-386.

275. Ruano G, Collins JM, Dorner AJ, Wang SJ, Guerciolini R, Huang SM. (2004) Pharmacogenomic data submissions to the FDA: clinical pharmacology case studies. Pharmacogenomics 5(5):513-517.

276. Rudberg I, Mohebi B, Hermann M, Refsum H, Molden E. (2008) Impact of the ultrarapid CYP2C19*17 allele on serum concentration of escitalopram in psychiatric patients. Clin Pharmacol Ther 83(2):322-327.

277. Ruhlen,M. A Guide to the World's Languages. Stanford: Stanford Univ Pr, USA; 1991.

278. Saadat M, Ansari-Lari M. (2007) Genetic polymorphism of glutathione S-transferase T1, M1 and asthma, a meta-analysis of the literature. Pak J Biol Sci 10(23):4183-4189.

279. Saag M, Balu R, Phillips E, et al. High sensitivity of human leukocyte antigen-B*5701 as a marker for immunologically confirmed abacavir hypersensitivity in white and black patients. Clin Infect Dis 2008;46:1111-1118.

280. Sabbagh A, Darlu P. (2006) SNP selection at the NAT2 locus for an accurate prediction of the acetylation phenotype. Genet Med 8(2):76-85.

281. Sabbagh A, Langaney A, Darlu P, Gerard N, Krishnamoorthy R, Poloni ES. (2008) Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet 9:21.

282. Sabeti PC. (2008) Natural selection: uncovering mechanisms of evolutionary adaptation to infectious disease. Nature Education 1(1):1.

283. Sadee W, Dai Z. (2005) Pharmacogenetics/genomics and personalized medicine. Hum Mol Genet 14 (2):R207-R214.

284. Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467.

285. Sangkuhl K, Berlin DS, Altman RB, Klein TE. (2008) PharmGKB: understanding the effects of individual genetic variants. Drug Metab Rev 40(4):539-551.

286. Scharf, S. (1995) PCR amplification of VNTRs. In: Innis,MA, Gelfand,DH, Sninsky,JJ, editors. PCR strategies. Academic Press pp. 161-175.

287. Scheuner MT, Sieverding P, Shekelle PG. (2008) Delivery of genomic medicine for common chronic adult diseases: a systematic review. JAMA 299(11):1320-1334.

288. Scordo MG, Aklillu E, Yasar U, Dahl ML, Spina E, Ingelman-Sundberg M. (2001) Genetic polymorphism of cytochrome P450 2C9 in a Caucasian and a black African population. Br J Clin Pharmacol. 52(4):447-50.

289. Scordo MG, Caputi AP, D'Arrigo C, Fava G, Spina E. (2004) Allele and genotype frequencies of CYP2C9, CYP2C19 and CYP2D6 in an Italian population. Pharmacol Res 50(2):195-200.

290. Seeringer A, Kirchheiner J. (2008) Pharmacogenetics-guided dose modifications of antidepressants. Clin Lab Med 28(4):619-626.

291. Service S, Sabatti C, Freimer N. (2007) Tag SNPs chosen from HapMap perform well in several population isolates. Genet Epidemiol 31(3):189-194.

292. Shen F, Huang J, Fitch KR, Truong VB, Kirby A, Chen W et al. (2008) Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. BMC Genet 9:27.

293. Shimada T, Yamazaki H, Mimura M, Inui Y, Guengerich FP. (1994) Interindividual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians. J Pharmacol Exp Ther 270(1):414-423.

294. Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C et al. (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112(4):387-399.

295. Sim SC, Ingelman-Sundberg M. (2006) The human cytochrome P450 Allele Nomenclature Committee Web site: submission criteria, procedures, and objectives. Methods Mol Biol 320:183-191.

296. Sim SC, Risinger C, Dahl ML, Aklillu E, Christensen M, Bertilsson L et al. (2006) A common novel CYP2C19 gene variant causes ultrarapid drug metabolism relevant for the drug response to proton pump inhibitors and antidepressants. Clin Pharmacol Ther 79(1):103-113.

297. Simon T, Verstuyft C, Mary-Krause M, Quteineh L, Drouet E, Meneveau N et al. (2009) Genetic determinants of response to clopidogrel and cardiovascular events. N Engl J Med 360(4):363-375.

298. Singer PA, Court EB, Bhatt A, Frew SE, Greenwood H, Persad DL et al. (2007) Applying genomics-related technologies for Africa's health needs. Afr J Med Med Sci 36 Suppl:7-14.

299. Sirima SB, Tiono AB, Gansane A, Diarra A, Ouedraogo A, Konate AT et al. (2009) The efficacy and safety of a new fixed-dose combination of amodiaquine and artesunate in young African children with acute uncomplicated Plasmodium falciparum. Malar J 8:48.

300. Sirugo G, Schim vdL, Sam O, Nyan O, Pinder M, Hill AV et al. (2004) A national DNA bank in The Gambia, West Africa, and genomic research in developing countries. Nat Genet 36(8):785-786.

301. Sistonen J, Fuselli S, Palo JU, Chauhan N, Padh H, Sajantila A. (2009) Pharmacogenetic variation at CYP2C9, CYP2C19, and CYP2D6 at global and microgeographic scales. Pharmacogenet Genomics 19(2):170-179.

302. Sistonen J, Sajantila A, Lao O, Corander J, Barbujani G, Fuselli S. (2007) CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. Pharmacogenet Genomics 17(2):93-101.

303. Slatkin M. (1993) Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264-279.

304. Slatkin M. (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-462.

305. Slone DH, Gallagher EP, Ramsdell HS, Rettie AE, Stapleton PL, Berlad LG et al. (1995) Human variability in hepatic glutathione S-transferase-mediated conjugation of aflatoxin B1-epoxide and other substrates. Pharmacogenetics 5(4):224-233.

306. Smith DA, Jones BC, Walker DK. (1996) Design of drugs involving the concepts and theories of drug metabolism and pharmacokinetics. Med Res Rev 16(3):243-266.

307. Smits KM, Gaspari L, Weijenberg MP, Dolzan V, Golka K, Roemer HC et al. (2003) Interaction between smoking, GSTM1 deletion and colorectal cancer: results from the GSEC study. Biomarkers 8(3-4):299-310.

308. Snyder, L. H. 2009. Inherited taste deficiency. Science 74, 151-152.

309. Snyder M, Weissman S, Gerstein, M (2009) Personal phenotypes to go with personal genomes Molecul Systems Biol 5:273.

310. Sokal RR, Rohlf FJ. Biometry. 2nd edition. San Francisco (CA): W.H. Freeman and Co.; 1981.

311. Srinivasan BS, Chen J, Cheng C, Conti D, Duan S, Fridley BL et al. (2009) Methods for analysis in pharmacogenomics: lessons from the Pharmacogenetics Research Network Analysis Group. Pharmacogenomics 10(2):243-251.

312. Stehle S, Kirchheiner J, Lazar A, Fuhr U. (2008) Pharmacogenetics of oral anticoagulants: a basis for dose individualization. Clin Pharmacokinet 47(9):565-594.

313. Steimer W, Zopf K, von AS, Pfeiffer H, Bachofer J, Popp J et al. (2005) Amitriptyline or not, that is the question: pharmacogenetic testing of CYP2D6 and CYP2C19 identifies patients with low or high risk for side effects in amitriptyline therapy. Clin Chem 51(2):376-385.

314. Steyn K, Fourie J, Bradshaw D. (1992) The impact of chronic diseases of lifestyle and their major risk factors on mortality in South Africa. S Afr Med J 82(4):227-231.

315. Stoehlmacher J, Park DJ, Zhang W, Groshen S, Tsao-Wei DD, Yu MC et al. (2002) Association between glutathione S-transferase P1, T1, and M1 genetic polymorphism and survival of patients with metastatic colorectal cancer. J Natl Cancer Inst 94(12):936-942.

316. Takahashi H, Wilkinson GR, Nutescu EA, Morita T, Ritchie MD, Scordo MG et al. (2006) Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese, Caucasians and African-Americans. Pharmacogenet Genomics 16(2):101-110.

317. Tan SH, Lee SC, Goh BC, Wong J. (2008) Pharmacogenetics in breast cancer therapy. Clin Cancer Res 14(24):8027-8041.

318. Tarlov AR, Brewer GJ, Carson PE, Alving AS. (1962) Primaquine sensitivity. Glucose-6-phosphate dehydrogenase deficiency: an inborn error of metabolism of medical and biological significance. Arch Intern Med 109:209-234.

319. Tate SK, Depondt C, Sisodiya SM, Cavalleri GL, Schorge S, Soranzo N et al. (2005) Genetic predictors of the maximum doses patients receive during clinical use of the anti-epileptic drugs carbamazepine and phenytoin. Proc Natl Acad Sci U S A 102(15):5507-5512.

320. Tayeb MT, Clark C, Ameyaw MM, Haites NE, Evans DA, Tariq M, Mobarek A, Ofori-Adjei D, McLeod HL. (2000) CYP3A4 promoter variant in Saudi, Ghanaian and Scottish Caucasian populations. Pharmacogenetics. 10(8):753-6.

321. Taylor AL, Ziesche S, Yancy C, Carson P, D'Agostino R, Jr., Ferdinand K et al. (2004) Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. N Engl J Med 351(20):2049-2057.

322. Templeton,AR. Population genetics and microevolutionary theory.John Wiley and Sons, Inc.; 2006.

323. Teufel A, Krupp M, Weinmann A, Galle PR. (2006) Current bioinformatics tools in genomic biomedical research (Review). Int J Mol Med 17(6):967-973.

324. The Uniprot Consortium. (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190-D195.

325. Thompson CA. (2007) FDA encourages genetics-aided warfarin dosing. Am J Health Syst Pharm 64(19):1994-1996.

326. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A et al. (2009) The genetic structure and history of Africans and African Americans. Science 324(5930):1035-1044.

327. Tishkoff SA, Verrelli BC. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293-340.

328. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM et al. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 40(5):623-630.

329. Tsai HJ, Shaikh N, Kho JY, Battle N, Naqvi M, Navarro D et al. (2006) Beta 2-adrenergic receptor polymorphisms: pharmacogenetic response

to bronchodilator among African American asthmatics. Hum Genet 119(5):547-557.

330. Tsao CC, Wester MR, Ghanayem B, Coulter SJ, Chanas B, Johnson EF et al. (2001) Identification of human CYP2C19 residues that confer S-mephenytoin 4'-hydroxylation activity to CYP2C9. Biochemistry 40(7):1937-1944.

331. Tsuchiya Y, Nakajima M, Takagi S, Taniya T, Yokoi T. (2006) MicroRNA regulates the expression of human cytochrome P450 1B1. Cancer Res 66(18):9090-9098.

332. United Nations World population Division. 2008.World Population Prospects. Population Division of the Department of Economic and Social Affairs. Electronic Citation.

333. Vandekerckhove L, Blot S, Vogelaers D. (2008) Abacavir hypersensitivity. N Engl J Med. 2008 358(23):2514-5.

334. van der Weide J, Steijns LS, van Weelden MJ, de HK. (2001) The effect of genetic polymorphism of cytochrome P450 CYP2C9 on phenytoin dose requirement. Pharmacogenetics 11(4):287-291.

335. Vansina, J. 1995.New Linguistic Evidence and ,The Bantu Expansion. Journal of African History 36.

336. van't Veer LJ, Bernards R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452(7187):564-570.

337. Vella A, Camilleri M. (2008) Pharmacogenetics: potential role in the treatment of diabetes and obesity. Expert Opin Pharmacother 9(7):1109-1119.

338. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. (2001) The sequence of the human genome. Science 291(5507):1304-1351.

339. Vladutiu GD. (2008) The FDA announces new drug labeling for pharmacogenetic testing: is personalized medicine becoming a reality? Mol Genet Metab 93(1):1-4.

340. Vogel F. Moderne probleme der humangenetik (1959). Ergebnisse der Innere Medizinsche und Kinderheilkunde, 12:52-62.

341. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L et al. (2002) Efficacy and safety of trastuzumab as a single agent in

first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 20(3):719-726.

342. Wang H, Tompkins LM. (2008) CYP2B6: new insights into a historically overlooked cytochrome P450 isozyme. Curr Drug Metab 9(7):598-610.

343. Wang J, Sonnerborg A, Rane A, Josephson F, Lundgren S, Stahle L et al. (2006) Identification of a novel specific CYP2B6 allele in Africans causing impaired metabolism of the HIV drug efavirenz. Pharmacogenet Genomics 16(3):191-198.

344. Wang JF, Wei DQ, Li L, Zheng SY, Li YX, Chou KC. (2007) 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. Biochem Biophys Res Commun 355(2):513-519.

345. Wang LL, Li Y, Zhou SF. (2009) A bioinformatics approach for the phenotype prediction of non-synonymous single nucleotide polymorphisms in human cytochrome P450s. Drug Metab Dispos 37(5):977-91.

346. Wang Z, Moult J. (2001) SNPs, protein structure, and disease. Hum Mutat 17(4):263-270.

347. Warren RM, Streicher EM, van Pittius NC, Marais BJ, van der Spuy GD, Victor TC et al. (2009) The clinical relevance of Mycobacterial pharmacogenetics. Tuberculosis (Edinb) 89(3):199-202.

348. Watson JD, Crick FH. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356):737-738.

349. Weckx S, De Rijk P, Van Broeckhoven C, Del Favero J. (2004) SNPbox: web-based high-throughput primer design from gene to genome. Nucleic Acids Res 32:W170-W172.

350. Weckx S, Del Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P et al. (2005) novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15(3):436-442.

351. Weinberg W. (1908) Ueber den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins fuer vaterlaendische Naturkunde in Wuerttemberg 64:368-382.

352. Weinshilboum RM, Wang L. (2006) Pharmacogenetics and pharmacogenomics: development, science, and translation. Annu Rev Genomics Hum Genet 7:223-245.

353. Weir BS, Cockerham CC. (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

354. Wennerholm A, Johansson I, Hidestrand M, Bertilsson L, Gustafsson LL, Ingelman-Sundberg M. (2001) Characterization of the CYP2D6*29 allele commonly present in a black Tanzanian population causing reduced catalytic activity. Pharmacogenetics 11(5):417-427.

355. Wennerholm A, Johansson I, Massele AY, Lande M, Alm C, Aden-Abdi Y, Dahl ML, Ingelman-Sundberg M, Bertilsson L, Gustafsson LL. (1999) Decreased capacity for debrisoquine metabolism among black Tanzanians: analyses of the CYP2D6 genotype and phenotype. Pharmacogenetics. 9(6):707-14.

356. WHO. (2004) World Health Report.

357. WHO. (2008) Report on the global AIDS epidemic.

358. WHO. 1-3-2005. Global tuberculosis control: surveillance, planning, financing.

359. Wigginton JE, Cutler DJ, Abecasis GR. (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76(5):887-893.

360. Wild CP, Yin F, Turner PC, Chemin I, Chapot B, Mendy M, Whittle H, Kirk GD, Hall AJ. (2000) Environmental and genetic determinants of aflatoxin-albumin adducts in the Gambia. Int J Cancer. 86(1):1-7.

361. Wilkinson, M. 6-19-2009.Biobanking on the future. Labtechnologist Newspaper.

362. Williams PA, Cosme J, Ward A, Angove HC, Matak VD, Jhoti H. (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature 424(6497):464-468.

363. Wojnowski L, Turner PC, Pedersen B, Hustert E, Brockmoller J, Mendy M et al. (2004) Increased levels of aflatoxin-albumin adducts are associated with CYP3A5 polymorphisms in The Gambia, West Africa. Pharmacogenetics 14(10):691-700.

364. Wright S. (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19:395-420.

365. Wu AC, Fuhlbrigge AL. (2008) Economic Evaluation of Pharmacogenetic Tests. Clin Pharmacol Ther 84(2):272-274.

366. Wu H, Dombrovsky L, Tempel W, Martin F, Loppnau P, Goodfellow GH et al. (2007) Structural basis of substrate-binding specificity of human arylamine N-acetyltransferases. J Biol Chem 282(41):30189-30197.

367. Wyen C, Hendra H, Vogel M, Hoffmann C, Knechten H, Brockmeyer NH et al. (2008) Impact of CYP2B6 983T>C polymorphism on non-nucleoside reverse transcriptase inhibitor plasma concentrations in HIV-infected patients. J Antimicrob Chemother 61(4):914-918.

368. Xie HG, Kim RB, Stein CM, Wilkinson GR, Wood AJ. (1999) Genetic polymorphism of (S)-mephenytoin 4'-hydroxylation in populations of African descent. Br J Clin Pharmacol 48(3):402-408.

369. Xing J, Witherspoon DJ, Watkins WS, Zhang Y, Tolpinrud W, Jorde LB. (2008) HapMap tagSNP transferability in multiple populations: general guidelines. Genomics 92(1):41-51.

370. Yasar U, Aklillu E, Canaparo R, Sandberg M, Sayi J, Roh HK, Wennerholm A. (2002) Analysis of CYP2C9*5 in Caucasian, Oriental and black-African populations. Eur J Clin Pharmacol. 58(8):555-8.

371. Yasar U, Eliasson E, Dahl ML, Johansson I, Ingelman-Sundberg M, Sjoqvist F. (1999) Validation of methods for CYP2C9 genotyping: frequencies of mutant alleles in a Swedish population. Biochem Biophys Res Commun 254(3):628-631.

372. Ye Z, Song H. (2005) Glutathione s-transferase polymorphisms (GSTM1, GSTP1 and GSTT1) and the risk of acute leukaemia: a systematic review and meta-analysis. Eur J Cancer 41(7):980-989.

373. Yeung S, Pongtavornpinyo W, Hastings IM, Mills AJ, White NJ. (2004) Antimalarial drug resistance, artemisinin-based combination therapy, and the contribution of modeling to elucidating policy choices. Am J Trop Med Hyg 71(2):179-186.

374. Yu A, Dong H, Lang D, Haining RL. (2001) Characterization of dextromethorphan O- and N-demethylation catalyzed by highly purified recombinant human CYP2D6. Drug Metab Dispos 29(11):1362-1365.

375. Yun CH, Yim SK, Kim DH, Ahn T. (2006) Functional expression of human cytochrome P450 enzymes in Escherichia coli. Curr Drug Metab 7(4):411-429.

376. Zang Y, Doll MA, Zhao S, States JC, Hein DW. (2007) Functional characterization of single-nucleotide polymorphisms and haplotypes of human N-acetyltransferase. Carcinogenesis 28(8):1665-1671.

377. Zanger UM, Klein K, Saussele T, Blievernicht J, Hofmann MH, Schwab M. (2007) Polymorphic CYP2B6: molecular mechanisms and emerging clinical significance. Pharmacogenomics 8(7):743-759.

378. Zanger UM, Turpeinen M, Klein K, Schwab M. (2008) Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. Anal Bioanal Chem 392(6):1093-1108.

379. Zhang F, Wang Y, Deng HW. (2008) Comparison of population-based association study methods correcting for population stratification. PLoS ONE 3(10):e3392.

380. Zhang WY, Tu YB, Haining RL, Yu AM. (2009) Expression and functional analysis of CYP2D6.24, CYP2D6.26, CYP2D6.27, and CYP2D7 isozymes. Drug Metab Dispos 37(1):1-4.

381. Zschocke J, Kohlmueller D, Quak E, Meissner T, Hoffmann GF, Mayatepek E. (1999) Mild trimethylaminuria caused by common variants in FMO3 gene. Lancet 354(9181):834-835.

382. Zupa A, Sgambato A, Bianchino G, Improta G, Grieco V, LA Torre G et al. (2009) GSTM1 and NAT2 Polymorphisms and Colon, Lung and Bladder Cancer Risk: A Case-control Study. Anticancer Res 29(5):1709-1714.

Appendix A1

No. of samples with successful genotypes for CYP2C9 re-sequencing study

NC_000010.9	cDNA	SNP	mRNA feature	effect	dbSNP	Hausa	Luo	Maasai	San	Shona	Venda	TZBantu	Total
96829291	-376	T>C	5'utr		rs9332103	12		11	13	8			44
96829916	251	T>C	intron		rs9332104	13	10	11	13	23	9	9	88
96833076	3411	T>C	intron		rs9332120	11	10	6	12	20	9	11	79
96833152	3487	A>G	intron		rs12769205	13	12	7	9	23	9	11	84
96833165	3499	T>A	intron		rs9332121		11			15	9	11	46
96838626	8963	T>C	intron		nrs		10			15	8	10	43
96838697	9032	G>C	intron		nrs	12	10	11	13	23	8	11	88
96838734	9069	G>A	intron		novel	12	11	11	13	23	8	10	88
96839116	9451	T>C	intron		rs17443251		11			15	8	9	43
96839976	10311	A>G	intron		rs9332129	12	6	10	13	22	8	8	79
96840012	10347	T>C	intron		novel		6			15	9	8	38
96840200	10535	A>G	exon 5	H251R (*9)	rs2256871	11	6	11	13	23	8	8	80
96840266	10601	wt>delA	exon 5	K273 fs (*6)	nrs	12		11	13	8			44
96863014	33349	A>G	intron		rs9332172	12	11	11	13	23	8	10	88
96863323	33658	A>G	intron		rs9332174	12	7	11	13	23	8	10	84
96872080	42415	C>T	intron		novel	13		9	12	4			38
96872134	42469	T>C	intron		rs9332197	13		11	11	7			42
96872184	42519	T>C	exon 7	I327T	novel	13		11	12	6			42
96872284	42619	G>C	exon 7	D360E (*5)	rs28371686	13	12	10	12	22	9	11	89
96877210	47545	A>T	intron		rs9332230		11			15	8	11	45
96877258	47593	T>C	intron		rs9332232	7	11	10	13	22	9	11	83
96877304	47639	C>T	intron		rs2298037		11			15	8	11	45
96879721	50056	A>T	intron		rs1934969	9	11	10	10	21	8	11	80
96879790	50125	C>T	intron		novel		11			15	8	11	45
96879861	50196	C>T	exon 9	A441A	rs2017319	13	11	11	13	23	9	11	91
96879959	50294	A>G	exon 9	N474S	novel		11			15	9	11	46
96879963	50298	A>T	exon 9	G475G	rs1057911		11			15	9	11	46
96880006	50341	G>T	exon 9	V490F	novel		11			15	8	11	45
96880078	50413	C>T	3'utr		rs9332240		11			15	8	11	45
96880099	50434	C>T	3'utr		rs9332241	13	11	11	13	23	8	11	90
96880168	50501	C>T	3'utr		rs9332243		11			15	8	11	45

No. of samples with successful genotypes for CYP2C19 re-sequencing study

NC_000010.9	cDNA	SNP	mRNA featu	effect	dbSNP	Hausa	Yoruba	Ibo	Luo	Maasai	Shona	Venda	TZ Bantu	Total
96653591	-97	T>C	5'utr		rs4986894	19	20	20	30	12	15	9	10	135
96653743	55	A>C	exon 1	I19L	rs17882687 (*19	19	20	20	30	13	15	9	10	136
96653787	99	T>C	exon 1	P33P	rs17885098	19	20	20	30	13	15	9	10	136
96653871	183	T>C	intron		rs17882201	19	20	20	30	13	14	9	10	135
96653876	188	G>A	intron		rs17881883	19	20	20	30	13	15	9	10	136
96653919	231	A>C	intron		novel	19	20	20	30	13	15	9	10	136
96665810	12122	A>G	intron		rs7916649	15	19	17	25	12	14	9	9	120
96665904	12306	G>A	intron		rs17878549	16	20	19	30	12	15	9	10	131
96666148	12460	G>C	exon 2	E92D	rs17878459	17	20	20	30	13	15	9	10	134
96666295	12607	wt>insC	intron		novel	17	20	20	30	13	15	9	10	134
96666325	12637	C>T	intron	splice site	novel	17	20	20	29	13	15	9	10	133
96666350	12662	A>G	intron	splice site	rs12769205	16	20	20	29	13	15	9	10	132
96666378	12690	G>A	exon 3	V113I	novel	15	19	20	29	13	15	9	10	130
96666472	12784	G>A	exon 3	R144H (*9)	rs17884712	17	20	20	29	13	15	9	10	133
96671557	17869	G>T	exon 4	R186P	novel	17	20	20	28	12	15	8	8	128
96671636	17948	G>A	exon 4	W212s (*3)	rs4986893	17	20	20	28	12	15	8	9	129
96671895	18207	G>A	intron		novel	17	20	20	28	12	15	8	8	128
96671917	18229	T>A	intron		rs17884938	16	20	20	28	12	15	8	8	127
96671942	18254	T>C	intron		novel	19	20	20	19	13	0	0	0	91
96672506	18818	T>C	intron	premiRNA	novel	19	20	20	29	13	15	9	10	135
96672599	18911	A>G	intron		rs7058784	29	20	20	29	13	15	9	10	145
96672764	19076	T>C	intron	splice site	novel	19	20	20	30	13	15	9	10	136
96672842	19154	G>A	exon 5	P227P (*2)	rs4244285	19	20	20	30	13	15	9	10	136
96673020	19332	G>A	intron	premiRNA	novel	19	20	20	28	13	14	9	9	132
96711141	57453	G>C	intron		novel	19	20	12	30	13	15	9	10	128
96711200	57512	A>G	intron		novel	19	20	12	30	13	14	9	10	127
96711255	57567	A>T	intron		novel	19	20	12	30	13	15	9	10	128
96711263	57575	T>C	intron		novel	19	20	12	29	13	15	9	10	127
96711325	57637	wt>delG	intron		novel	19	20	12	29	13	15	9	10	127
96711366	57678	T>G	intron		rs28399511	19	20	12	28	12	15	9	10	125
96711428	57740	G>C	intron		rs4417205	19	20	12	30	13	15	9	10	128
96711877	57969	G>C	intron		novel	17	19	12	28	12	15	9	10	122
96733848	80160	C>T	exon 7	V330V	rs3758580	17	19	12	28	12	12	9	10	119
96733849	80161	G>A	exon 7	V331I	rs3758581	17	19	12	28	12	12	9	10	119
96734317	80629	T>A	intron		novel	17	11	8	26	11	10	9	10	102

96740794	87106	T>C	intron		rs4917623	20	20	12	29	13	15	9	10	128
96740978	87290	T>C	exon 8	R410C (*13)	rs17879685	20	20	12	30	13	15	9	10	129
96741001	87313	A>C	exon 8		rs17886522	20	20	12	30	13	15	9	10	129
96741119	87422	A>G	intron		novel	20	20	12	30	13	15	9	10	129
96741183	87475	G>C	intron		rs17880188	20	20	12	29	13	14	9	10	127
96741210	87522	C>T	intron		rs17885587	20	20	12	29	13	15	9	10	128
96743266	89578	T>A	intron		rs12779363	16	18	9	27	12	15	9	10	116
96743597	89909	C>T	intron		rs12268020	16	18	9	27	12	15	9	10	116
96743699	90011	A>G	intron	splice site	rs4451645	16	18	9	27	12	15	9	10	116
96743897	90209	A>C	exon 9	X491C; 26 extra a	nrs	18	18	12	28	13	15	9	10	123
96743989	90301	C>T	3'utr		novel	19	18	12	28	13	15	9	10	124
96743990	90302	C>T	3'utr		novel	19	18	12	28	13	15	9	10	124
96744221	90533	C>T	3'utr		novel	17	17	10	26	13	12	9	9	113

No. of samples with successful genotypes for CYP2D6 re-sequencing study

M33388	cDNA	SNP	mRNA featu	effect	db SNP	Hausa	Yoruba	Ibo	Luo	Maasai	Shona	Venda	TZBantu	Total
1444	-175	G>A	5'utr		rs1080993	11	17	18	27	11	14	8	10	116
1469	-150	C>T	5'utr		nrs	11	17	19	28	11	15	8	10	119
1534	-85	T>C	5'utr		nrs	13	17	19	28	11	15	8	10	121
1577	-42	wt>insG	5'utr		rs28371695	13	17	19	28	11	15	8	10	121
1696	77	G>A	exon 1	R26H (*43)	rs28371696	12	17	19	28	10	15	8	10	119
1701	82	C>T	exon 1	R28C (*22)	nrs	14	17	19	28	10	15	8	10	121
1719	100	C>T	exon 1	P34S (*10)	rs1065852	13	17	20	28	10	14	8	10	120
1833	214	G>C	intron		rs1080995	4	11	15	21	10	7	6	7	81
1840	221	C>A	intron		rs1080996	4	11	15	20	10	7	5	7	79
1842	223	C>G	intron		rs1080997	4	11	15	21	10	6	5	6	82
1846	227	T>C	intron		rs1080998	4	11	15	21	10	7	4	5	77
1851	232	G>C	intron		rs1080999	4	11	15	21	10	5	4	4	74
1852	233	A>C	intron		rs1080999	4	11	15	20	10	6	4	5	75
1864	245	A>G	intron		rs1081000	4	11	15	18	10	5	4	3	70
1929	310	G>T	intron		rs28371699	4	11	15	16	8	2	4		60
2273	654	C>T	intron		novel		2	4	14		12	7	7	46
2365	746	C>G	intron		nrs				11		15	8	10	44
2462	843	T>G	intron		rs28371702	7	15	15	13	12	15	8	10	95
2625	1006	C>T	exon 2	R101R	novel	5	15	16	23	12	15	9	10	105
2642	1023	C>T	exon 2	T107I (*17)	rs28371706	5	15	16	23	12	15	8	10	104
2658	1039	C>T	exon 2	F112F	rs1081003	4	15	16	23	12	15	8	10	103
2686	1067	T>G	intron	splice site	novel	4	15	16	23	12	15	5	10	100
3227	1608	G>A	exon 3	V119M	novel	19	20	20	28	12	15	9	10	133
3240	1621	G>T	exon 3	R123L	novel	19	20	20	28	12	15	9	10	133
3278	1659	G>A	exon 3	V136M (*29)	rs1058164	19	20	20	27	12	15	9	10	132
3280	1661	G>C	exon 3	V136V	rs28371706	19	20	20	27	12	15	9	10	132
3335	1716	G>A	exon 3	E155K (*45)	rs28371710	19	20	20	28	12	15	9	10	133
3465	1846	G>A	intron	182 splicing defe	nrs	19	20	20	28	12	15	9	10	133
3483	1864	wt>delGT	exon 4	ins 3 aa	nrs	19	20	20	28	12	15	9	10	133
3485	1866	C>T	exon 4	N175N	nrs	19	20	20	28	12	15	9	10	133
3488	1869	T>C	exon 4	G176G	nrs	19	20	20	28	12	15	9	10	133
3617	1998	T>C	exon 4	F219F	novel			1	9		15	9	10	44
4194	2575	C>A	exon 5	P267P	nrs				11		15	9	10	45
4221	2602	G>T	exon 5	L276L	novel				11		15	9	10	45
4280	2661	G>A	intron		nrs				11		15	9	10	45

4379	2760	T>A	intron		novel				11		15	9	10	45
4469	2850	C>T	exon 6	R296C	nrs				11		15	9	10	45
4607	2988	G>A	intron	splicing defect	nrs				11		15	9	10	45
4802	3183	G>A	exon 7	V338M (*29)	nrs	16	20	19	27	12	15	8	8	125
4873	3254	T>C	exon 7	H361H	rs2743457	16	20	20	27	12	15	8	8	126
4880	3259	wt>insTG	exon 7	375 fs (*42)	nrs	16	20	20	27	12	15	8	8	126
5003	3384	A>C	intron		nrs	15	20	19	27	13	15	8	8	125
5016	3397	C>A	intron		novel	15	19	19	27	13	15	8	8	124
5180	3561	G>C	intron		novel	112	19	19	27	11	14	8	8	218
5201	3582	A>G	intron		nrs	12	19	19	27	12	15	8	8	120
5203	3584	G>A	intron		nrs	12	19	19	27	12	15	8	8	120
5328	3707	G>A	intron		nrs	13	19	19	26	13	15	7	8	120
5349	3721	wt>delGT	intron		nrs	12	19	19	26	13	15	8	8	120
5409	3790	C>T	intron	splice site	nrs	15	19	19	26	13	15	8	8	123
5472	3853	G>A	exon 8	E410K (*27)	nrs	16	19	19	26	12	15	8	8	123
5652	4033	C>T	intron	splice site	novel	17	19	20	27	12	15	8	8	126
5676	4057	G>A	exon 9	G445E	novel	18	19	19	27	12	15	8	8	126
5799	4180	G>C	exon 9	S486T	rs1135850	17	20	19	27	12	15	8	9	127
6013	4394	wt>delAG	3'utr		novel	17	20	19	27	12	15	8	9	127
6020	4401	C>T	3'utr		nrs	17	20	19	27	12	14	8	9	126
6100	4481	G>A	3'utr		nrs	17	20	20	27	12	15	7	8	126
6275	4656	wt>delACA	3'utr		nrs	9	10	18	20	8	6	6	6	89
6341	4722	T>G	3'utr		nrs	4	14	13	14	1	2	1	2	51

No. of samples with successful genotypes for NAT2 re-sequencing study

NC_000008.9	cDNA	SNP	effect	db SNP	Hausa	Yoruba	Ibo	Luo	Maasai	San	Total
8950	191	G>A	R64Q (*14)	rs1801279	20	18	19	15	12	40	124
9041	282	C>T	Y94Y	rs1041983	20	18	19	16	12	40	125
9100	341	T>C	I114T (*5)	rs1801280	20	18	19	15	12	40	124
9162	403	C>G	L135V	nrs	20	18	19	15	12	40	124
9231	472	A>C	I158L	novel	20	18	19	15	12	40	124
9240	481	C>T	L161L	rs1799929	20	18	19	15	12	40	124
9346	589	C>T	R197X	novel	20	18	19	16	12	40	125
9349	590	G>A	R197Q (*6)	rs1799930	19	18	19	15	12	38	121
9400	641	C>T	T214I	novel	20	18	19	16	12	40	125
9442	683	C>T	P226L	nrs	20	18	19	16	12	40	125
9525	766	A>G	K256E	nrs	20	18	19	16	12	40	125
9562	803	A>G	K268R	rs1208	19	18	19	16	12	40	124
9568	809	T>C	I270T	novel	19	18	19	16	12	40	124
9597	838	G>A	V280M	nrs	18	18	19	16	12	40	123
9616	857	G>A	G286E (*7)	rs1799931	18	18	19	16	12	40	123

No. of samples with successful genotypes for FMO3 Taqman assay

	Luo	Maasai	Kikuyu	Shona	Venda	San	Hausa	Yoruba
D132H	99	143	97	99	63	63	100	100
E158K	99	143	97	99	63	63	100	100
V257M	99	143	97	99	63	63	100	100
E308G	99	143	97	99	63	63	100	100
L360P	99	143	97	99	63	63	100	100

No. of samples with successful genotypes for Sequenom assay study

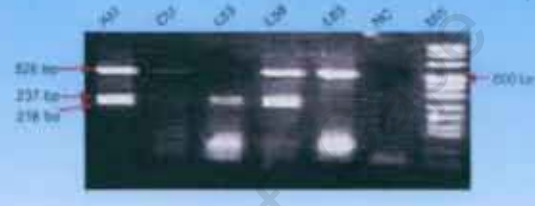
Assay	SNP	Hausa	San	Shona	Maasai
1030	NAT2rs4646247	46	40	45	50
1031	NAT2rs1801280	46	39	46	50
1033	NAT2rs721399	46	40	45	50
1034	NAT2rs7832071	45	40	45	50
1039	NAT2rs4646246	46	40	45	50
1041	NAT2rs1799930	47	40	46	50
1043	NAT2rs1799931	47	40	45	49
1052	NAT2rs721398	48	40	45	49
1060	NAT2rs1799929	48	40	45	50
1062	NAT2rs1208	47	39	45	49
1065	NAT2rs1801279	47	40	45	50
1032	CYP2B6rs3786547	47	40	45	49
1042	CYP2B6rs8192712	47	40	46	49
1044	CYP2B6rs3745274	48	40	45	51
1045	CYP2B6rs7259965	44	33	39	46
1046	CYP2B6rs2308606	48	40	45	48
1047	CYP2B6rs8100458	47	40	45	50
1048	CYP2B6rs12721649	48	40	45	49
1053	CYP2B6rs10426235	48	40	45	49
1059	CYP2B6rs2279343	48	40	45	51
1064	CYP2B6rs8192711	47	40	46	49
1035	CYP2C19rs4986894	47	40	45	50
1036	CYP2C19rs4244285	48	40	45	49
1037	CYP2C19rs12779363	47	40	46	49
1038	CYP2C19rs4304697	47	40	45	48
1040	CYP2C19rs4986893	44	37	44	43
1056	CYP2C19rs4388808	46	38	45	47
1057	CYP2C19rs7088784	46	38	45	47
1061	CYP2C19rs11597626	46	38	46	47
1063	CYP2C19rs10509676	39	37	40	47

No. of samples with successful genotypes for RFLP analysis study

Gene	Allele name	Tanzanian	Shona	Venda	Kikuyu	Luo	Maasai	Igbo	Yoruba	Hausa	San	Total
CYP2C9	*2		68								55	123
	*3		65								53	118
CYP2C19	*2	71	73		92	92	132	101	101	97	60	819
	*3	60										60
CYP2D6	*2		73		50		57				59	239
	*4		71		39	49	33	30	29	26	63	340
	*17		74		47	74	73	68	76	49	63	524
	*29		77		46	43	42	45			77	330
NAT2	*5				70	84	115	93	35	98	45	540
	*6					86	116	102	100	94	42	540
	*7							55				55
GST	M1 del/del				89	29	147	103	100	99	45	612
	T1 del/del				89	29	146	103	100	99		586
CYP2B6	*6	70	79	42	84	89	128	89	81	82	60	804

Appendix A2

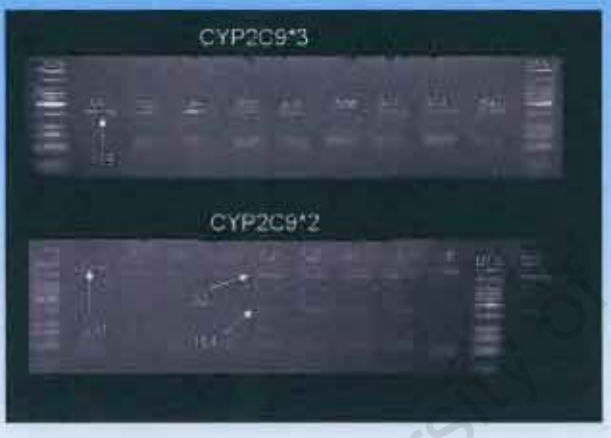
CYP2B6*6



Variant SNP sequence not cut with BstUI; genotype confirmed with sequenom assay
AM +/-, CM inc, L53 -/-, L59 +/-, L63 +/+

+/+ homozygous for test allele
+/- heterozygous for test allele
-/- homozygous for wild-type allele
inc inconclusive
NC no-DNA control
MW molecular weight marker

CYP2C9

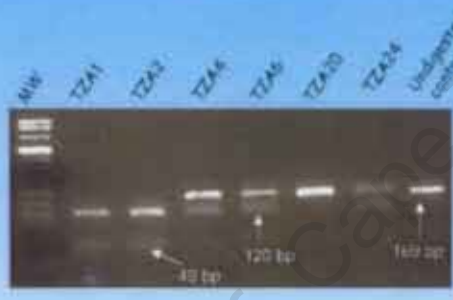


CYP2C9*3 (1075 G>A). Variant SNP sequence cut with KpnI restriction enzyme; genotypes confirmed by sequencing.
S28, S30, S33, S35, S36, S37, S39, S41 -/-

CYP2C9*2 (430 C>T). Variant SNP sequence not cut with AvaII restriction enzyme: genotypes confirmed by sequencing
S16, S18, S19, S21, S23, S2 -/-

+/+ homozygous for test allele
+/- heterozygous for test allele
-/- homozygous for wild-type allele
inc inconclusive
UC undigested control
LMW molecular weight marker

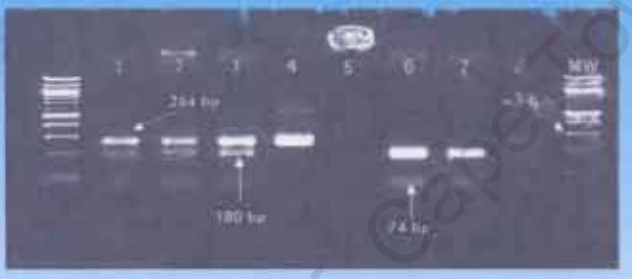
CYP2C19*2



Variant SNP sequence not cut with SmaI restriction enzyme, genotypes were confirmed by sequencing.
TZA1 -/-, TZA2 -/-, TZA4 +/-, TZA6 +/-, TZA20 +/+, TZA24 inc

+/+ homozygous for test allele
+/- heterozygous for test allele
-/- homozygous for wild-type allele
inc inconclusive
MW molecular weight marker

CYP2D6 *17



Variant SNP sequence not cut with FokI restriction enzyme; genotypes confirmed by sequencing
Lane 1, 2, 3 +/-; Lane 4 +/+; Lane 5,8 inc; Lane 6, 7 -/-

+/+ homozygous for test allele
+/- heterozygous for test allele
-/- homozygous for wild-type allele
inc inconclusive
MW molecular weight marker

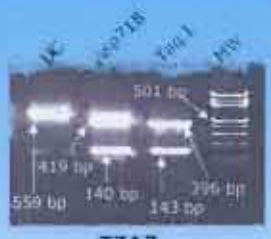
GSTM1

Homozygous deletion of GST gene is detected by absence of 273 bp PCR product. Beta actin amplification of 600bp fragment was used as internal control for DNA amplification.

TZA1 -, TZA2 -, TZA3 -, TZA4 +, TZA5 -, TZA6 -

\+ homozygous deletion for GST
\- at least 1 copy of GSTM1 present
MW molecular weight marker

NAT2*5 and *6



TZA7

TZA8

TZA10

NAT2*5 (341 T>C)	*Asp718*	419 +140bp (wt) 559bp (mt)	**TZA7, TZA8, TZA9, TZA10 +/-**
NAT2*6 (590 G>A)	*TaqI*	20+143+169+227bp (wt) 396bp (mt)	**TZA7 +/+, TZA8 -/-, TZA10 +/-**

TZA9

For each sample a 559 PCR fragment was tested for digestion with Asp 718, TaqI, BamHI in 3 separate reactions
Digestion patterns are as shown, genotypes were confirmed by sequencing

+/+ homozygous for test allele
+/- heterozygous for test allele
-/- homozygous for wild-type allele
UC undigested control
MW molecular weight marker

Appendix B

Geographical location of sampled populations