

https://dspace.org/

Genetic polymorphisms in xenobiotic (or drug) metabolizing enzyme genes among 18 sub-Saharan African populations: a window into genetic diversity

Makkan, Heeran

2014-06
http://hdl.handle.net/10539/15432

Downloaded from a DSpace repository. https://dspace.org/

Genetic polymorphisms in xenobiotic (or drug) metabolizing enzyme genes among 18 sub-Saharan African populations: a window into genetic diversity

HEERAN MAKKAN

A dissertation submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Masters of Science in Medicine 2014

DECLARATION

I, Heeran Makkan declare that this dissertation is my own unaided work. It is being submitted for the Degree of Master of Science in Human Genetics at the University of the Witwatersrand, Johannesburg. It has not been submitted for any other degree at this or any other university.

_____ _____

Heeran Makkan Date

ABSTRACT

Many loci coding for xenobiotic metabolising enzymes, especially those involved in carcinogen metabolism, confer susceptibility to various types of cancers. These genes have been poorly investigated in sub-Saharan African populations, where the genetic variation that exists is relatively unknown. The primary objectives of the study are to determine the frequency variation among 15 loci in sub-Saharan Africans, the level of genetic diversity, and the genetic affinities among sub-Saharan Africans. Secondary, the study aims to evaluate the implication of these variants in disease susceptibility, especially cancer.

The study population comprised of 1880 unrelated individuals from 18 sub-Saharan African populations. DNA samples were used to examine genetic variation for phase I metabolism genes *CYP1A1, CYP1A2, CYP2A6, CYP2D6, CYP2E1*, and phase II metabolism genes *GSTM1, GSTT1, GSTP1* and *NAT2*. A single base extension (SBE) method was designed and used to genotype single nucleotide polymorphisms (SNP): CYP1A1*2A and *2C; CYP1A2*1C and *1F; CYP2A6*7 and *8; CYP2D6*3A (2549delA) and CYP2D6*4(1846G>A); CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I); GSTP1*Ile105Val and *Ala114Val; and NAT2*14A. To investigate the presence of null mutations GSTM1*0 and GSTT1*0 a previously reported multiplex PCR method was used. The distribution of mutations in the sample was interpreted and compared with data from literature respectively.

Mutations CYP1A1*2A, CYP1A2*1C, CYP1A2*1F, CYP2A6*7, CYP2D6*4 and GSTP1*Ile105Val mutations was found in most sub-Saharan Africans, while CYP1A1*2C, CYP2A6*8, CYP2D6*3A and GSTP1*Ala114Val mutations were almost non-existent. Both GSTM1*0 and GSTT1*0 mutations were present in all populations, with GSTT1*0 most frequent. The distribution of NAT2*14A confirms previous reports of its exclusive existence in Africans. Hardy-Weinberg Equilibrium (HWE) and Tajima's D statistic tests showed none of the mutations were under selection. The genetic affinities of sub-Saharans were analysed. Bantu-speakers were closely related with little correlation to their geographic locations. Khoisan-speakers were closely related, genetically most distinct and oldest among populations. Pygmies were similarly distinct from most populations and one of the oldest surviving populations. The data further supports previous reports that the Khwe are descendants of an east African pastoralist group. AMOVA analyses revealed language as a major confounder among sub-Saharans. Haplotypes were inferred to determine their distribution and to understand their significance in populations with respect to their functional relevance.

The study has confirmed previous reports of genetic histories of these sub-Saharan African populations. In unravelling the distribution of these mutations, the study has added to the global picture of these mutations. In doing so, the data may add value to the design of future cancer studies and pharmacological studies. The study also highlights the importance of elucidating ancestral relations of populations, more specifically linguistic and anthropological relationships, and to include in the design of future clinical trials in Africa.

DEDICATION

In loving memory of my grandmother,

Manibhen ‘Ba’ Ghosai Wallabh

ACKNOWLEDGEMENTS

I am extremely grateful to all individuals who voluntarily provided samples towards the project.

I would like to acknowledge the National Health Laboratory Service, University of the Witwatersrand, the National Research Fund and the Medical Research Council for funding this study.

To my colleagues at the NHLS, I would like to thank you for the assistance in sample collection and DNA isolated for the sample bank of the Human Genomic Diversity and Disease Research Unit. Special thanks to Dr. Carina M. Schlebusch and Thijessen Naidoo for their guidance towards the design of the SBE protocol and analysis methods used for the project.

My sincerest thank you to Prof. Himla Soodyall for her mentorship and support throughout the course of my career. The lessons learnt have been tremendous in my development as a researcher.

To Prof. Arnold L. Christianson, thank you for the support and guidance during my employment with the NHLS, Division of Human Genetics.

Thank you to Prof. Trevor Jenkins for your guidance and the thought-provoking chats we have shared.

To my colleagues at the Aurum Institute, thank you. Especially, Dr. Salome Charalambous and Dr. Vinodh A. Edward for their understanding, support and affording me the time to complete this study.

To my family and friends, thank you for the kind support and love you give me.

To my parents Ashwin and Bhanoo, I am eternally grateful for the never-ending love and support you have shown me, without whom my successes have no meaning.

To my loving wife Candice - thank you for the constant support, the constant encouragement to complete the study, and the enormous care and love you give me.

TABLE OF CONTENTS

DECLARATION ii

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

LIST OF FIGURES 5

LIST OF TABLES 7

LIST OF ABBREVIATIONS 8

CHAPTER 1 INTRODUCTION 10

1.1 The sub-Saharan African population history 12

1.1.1 South Africa 14

1.1.2 Botswana 19

1.1.3 Mozambique 19

1.1.4 Namibia 20

1.1.5 Zambia 21

1.1.6 Uganda 22

1.1.7 Democratic Republic of Congo 22

1.1.8 Central African Republic 22

1.2 Genetic studies in human diseases 24

1.3 Cancer epidemic in Africa 25

1.4 Interaction of metabolic genes in cancer progression 31

1.5 Drug metabolising enzymes (DME) 33

1.6 Genes under investigation 34

1.6.1 Phase I genes 34

1.6.1.1 *CYP1A1* gene 35

1.6.1.2 *CYP1A2* gene 36

1.6.1.3 *CYP2A6* gene 37

1.6.1.4 *CYP2D6* gene 38
1.6.1.5 *CYP2E1* gene 40
1.6.2 Phase II genes 42
1.6.2.1 Glutathione-S-transferase genes 42
1.6.2.2 N-acetylation genes 45
1.7 Study objectives 46
CHAPTER 2 SUBJECTS, MATERIALS AND METHODS 47
2.1 Subjects 47
2.2 Comparative data 50
2.3 Materials and methods 50
2.3.1 DNA extractions and quantification 50
2.3.2 Null allele typing methods 50
2.3.3 Detection of SNP markers using the Single Base Extension method 52
2.3.3.1 Primer design 53
2.3.3.2 Multiplex PCR optimization 54
2.3.3.3 Genotyping of SNP markers using the SBE method 58
2.4 Statistical analysis 60
2.4.1 Population statistics and neutrality testing 61
2.4.1.1 Hardy-Weinberg equilibrium 61
2.4.1.2 Neutrality testing 61
2.4.1.3 Population diversity 62
2.4.2 Population affinity 62
2.4.3 Population structure 63
CHAPTER 3 RESULTS 65
3.1. Mutation detection 65
3.2. Phase I metabolizing genes 69
3.2.1. CYP1A1*2A allele frequencies 69
3.2.2. CYP1A1*2C allele frequencies 70

3.2.3. CYP1A2*1C allele frequencies 71
3.2.4. CYP1A2*1F allele frequencies 77
3.2.5. CYP2A6*7 allele frequencies 77
3.2.6. CYP2A6*8 allele frequencies 78
3.2.7. CYP2D6*3A (2549delA) allele frequencies 78
3.2.8. CYP2D6*4 (1846G>A) allele frequencies 84
3.2.9. CYP2E1*5B(*Pst*I) allele frequencies 86
3.2.10. CYP2E1*5B(*Rsa*I) allele frequencies 86
3.3. Phase II metabolizing genes 87
3.3.1. GSTM1*0 genotype frequencies 90
3.3.2. GSTT1*0 genotype frequencies 90
3.3.3. GSTP1*Ile105Val allele frequencies 96
3.3.4. GSTP1*Ala114Val allele frequencies 98
3.3.5. NAT2*14A allele frequencies 98
3.4. Hardy-Weinberg Equilibrium and Tajima's D statistic tests performed 102
3.5. Average gene diversity and average heterozygosity 113
3.6. Population affinities and evolution measured by genetic distance 116
3.7. Population structure 123
CHAPTER 4 DISCUSSION 125
4.1. Phase I metabolizers 125
4.2. Phase II metabolizers 132
4.3. Role of Phase I and Phase II metabolisers in Africa 136
4.4. Population genetic diversity 138
4.5. Genetic relationships of sub-Saharan Africans 139
4.6. Limitations of the study 143
4.7. Future studies and concluding remarks 144
REFERENCES 146
APPENDICES 189

Appendix A: Ethics clearance certificates.....189

Appendix B: Comparative data.....192

Appendix C: SNP Marker regions with DNA Sequence, PCR primer and SBE primer sequence.....201

Appendix D: Pair-wise genetic distance (F_{st}) for indigenous populations.....208

LIST OF FIGURES

Figure 1 Methods of incidence data estimation for GLOBOCAN 2008 26
Figure 2 Age-adjusted DALY's of cancer in 2008 worldwide 28
Figure 3 Most common cancer sites in Africa by sex and country are shown, GLOBOCAN 2008 30
Figure 4 Pathway from carcinogen exposure to cancer development 33
Figure 5 Map indicating position where samples were collected 48
Figure 6 Examples of electropherogram results obtained during genotyping of SNP markers using SBE technique. 67
Figure 7 Gel picture showing results of GSTM1 and GSTT1 multiplex reaction 68
Figure 8 Bar chart showing CYP1A1*2A allele proportions worldwide 75
Figure 9 Bar chart showing CYP1A1*2C allele proportions worldwide 76
Figure 10 Bar chart showing CYP1A2*1C allele proportions worldwide 79
Figure 11 Bar chart showing CYP1A2*1F allele proportions worldwide 80
Figure 12 Bar chart showing CYP2A6*7 allele proportions worldwide 81
Figure 13 Bar chart showing CYP2A6*8 allele proportions worldwide 82
Figure 14 Bar chart showing CYP2D6*3A allele proportions worldwide 83
Figure 15 Bar chart showing CYP2D6*4 allele proportions worldwide 85
Figure 16 Bar chart showing CYP2E1*5B(*Pst*I) allele proportions worldwide 88
Figure 17 Bar chart showing CYP2E1*5B(*Rsa*I) allele proportions worldwide 89
Figure 18 Bar chart showing GSTM1*0 genotype proportions worldwide 93
Figure 19 Bar chart showing GSTT1*0 genotype proportions worldwide 94
Figure 20 Bar chart showing the proportion of individuals who are homozygous for both GSTM1*0 and GSTT1*0 mutations worldwide 95
Figure 21 Bar chart showing GSTP1*Ile105Val allele proportions worldwide 97
Figure 22 Bar chart showing GSTP1*Ala114Val allele proportions worldwide 100
Figure 23 Bar chart showing NAT2*14A allele proportions worldwide 101
Figure 24 Average gene diversity across 13 SNP loci in study populations 114
Figure 25 Average observed heterozygosity across 13 SNP loci in study population 115
Figure 26 Neighbour-Joining population tree for indigenous populations 117
Figure 27 Principal component analysis of Fst values from allele distribution of 13 loci – component 1 versus component 2 120
Figure 28 Principal component analysis of Fst values from allele distribution of 13 loci – component 1 versus component 3 121

Figure 29 Results of the Structure runs (K=2 to K=6) for SNP loci in genes *CYP1A1, CYP1A2, CYP2A6, CYP2D6, CYP2E1, GSTP1* and *NAT2* 124

LIST OF TABLES

Table 1 Study population-sample origin, ethnicity, language, population code and sample size 49
Table 2 Primer name and sequences for GSTM1 and GSTT1 null allele detection 51
Table 3 List of SNP markers typed using SBE reaction along with primer sequences, fragment lengths, and alleles colour coded to the respective ddNTP fluorescent dye contained in the SBE reaction mix 56
Table 4 Multiplex PCR primer and SBE primer concentrations used in SBE method . 59
Table 5 SNP loci mutant allele frequencies only – current study population 72
Table 6 GSTM1 and GSTT1 null genotype frequency..... 92
Table 7 HWE and Tajima's D statistic tests performed for JU population..... 103
Table 8 HWE and Tajima's D statistic tests performed for XUN population 103
Table 9 HWE and Tajima's D statistic tests performed for KHWE population 104
Table 10 HWE and Tajima's D statistic tests performed for NAMA population 104
Table 11 HWE and Tajima's D statistic tests performed for SEB population..... 105
Table 12 HWE and Tajima's D statistic tests performed for MOZ population..... 105
Table 13 HWE and Tajima's D statistic tests performed for HER population 106
Table 14 HWE and Tajima's D statistic tests performed for HIM population..... 106
Table 15 HWE and Tajima's D statistic tests performed for OVA population 107
Table 16 HWE and Tajima's D statistic tests performed for DAMA population 107
Table 17 HWE and Tajima's D statistic tests performed for UGA population 108
Table 18 HWE and Tajima's D statistic tests performed for ZAM population 108
Table 19 HWE and Tajima's D statistic tests performed for DRC population..... 109
Table 20 HWE and Tajima's D statistic tests performed for CAR_U population 109
Table 21 HWE and Tajima's D statistic tests performed for CAR_P population 110
Table 22 HWE and Tajima's D statistic tests performed for SAC population 110
Table 23 HWE and Tajima's D statistic tests performed for SAW population 111
Table 24 HWE and Tajima's D statistic tests performed for SAI population 111
Table 25 Exact Test of sample differentiation based on allele frequencies 112
Table 26 AMOVA analysis summary of indigenous African sample 122

LIST OF ABBREVIATIONS

%	Percentage
°C	Degrees Celsius
µl	microlitre
µM	micromolar
3'	Three prime
5'	Five prime
A	Adenine
ABI	Applied Biosystems
AMOVA	analysis of molecular variance
bp	Base pairs
C	Cytosine
CAR_P	Central African Republic Pygmy
CAR_U	Central African Republic Ubangian-speakers
CYP	Cytochrome P450
DAMA	Dama
dATP	Deoxyadenosine-5'-triphosphate
dCTP	Deoxycytidine-5'-triphosphate
del	Deletion
ddH2O	Deionised distilled water
ddNTPs	Dideoxyribonucleotide triphosphate
ddTTP	Dideoxyribothymidine triphosphate
dGTP	Deoxyguanosine-5'-triphosphate
DME	Drug Metabolising Enzyme
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DRC	Democratic Republic of Congo Bantu-speakers
dTTP	Deoxythymidine-5'-triphosphate
EDTA	Ethylenediaminetetra-acetic acid
F	Forward primer
G	Guanine
GST	Glutathione-S-transferase
HER	Herero
HIM	Himba
HPLC	high performance liquid chromatography

HWE	Hardy-Weinberg Equilibrium
JU	Ju|'hoansi
kb	Kilobase
KHWE	Khwe
M	Molar
MgCl2	Magnesium Chloride
min	Minute
ml	Millilitre
mM	Millimolar
MOZ	Mozambican Bantu-speakers
mtDNA	mitochondrial DNA
NaCl	Sodium chloride
NAMA	Nama
NAT	N-acetylation
ng	Nanograms
NHLS	National Heath Laboratory Service
OVA	Ovambo
P	Probability
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
R	Reverse primer
RFLP	restriction fragment length polymorphism
SAC	South African Coloureds
SAI	South African Indians
SAW	South African Whites
SBE	Single base extension
SEB	South African Bantu-speakers
sec	Second
T	Thymine
TBE	Tris-Boric acid-EDTA
Taq	Thermus aquaticus
U	Units of enzyme microgram per microlitre
UGA	Ugandan Bantu-speakers
WITS	University of the Witwatersrand
XUN	!Xun
ZAM	Zambian Bantu-speakers

CHAPTER 1 INTRODUCTION

Understanding the genetic variability of global populations has been the focus of numerous population and evolutionary genetic studies (Cavalli-Sfroza et al., 1994; Jorde et al., 1998; Lambert and Tishkoff, 2009; Campbell and Tishkoff 2010; Schlebusch et al., 2012). Cavalli-Sfroza et al. (1994) used classical polymorphisms to understand variation among global populations. Recent research have unravelled the genetic variation that exists on the maternally inherited mitochondrial DNA (mtDNA), the paternally inherited non-recombining region of the Y sex-chromosome (NRY) and autosomal DNA (Jenkins, 1986; Tishkoff et al., 2007; Behar et al., 2008; Naidoo et al., 2010; Schlebusch, 2010a; Schlebusch et al., 2012; Badro et al., 2013; Montano et al., 2013). These findings contributed to widely recognised and consolidated history of human evolution. Modern humans evolved about 200,000 years ago and have lived the longest in Africa. Human migration took place across North-East Africa through the Arabian Peninsula and beyond into Asia, Europe and finally the Americas. The dispersion of modern humans outside Africa occurred over different time periods spanning 80,000 to 15,000 years ago (Campbell and Tishkoff 2010). This resulted in modern humans emerging today with novel genetic and phenotypic adaptations where the genetic variation found outside Africa is a subset thereof that exists in African populations. The evolution of humans resulted in large global demographic variation, where contemporary populations vary in thousands of different languages, ethnicities and cultures (Campbell and Tishkoff 2010).

The demographic history of Africans has been shaped by many fluctuations in population size, migration events, admixture and extensive population structure. This process would have occurred in contrasting environments where Africans lived

in tropical rainforests, swamps, savannas, deserts and mountain highlands. In adapting to these niches, African populations have developed a variety of subsistence methods in hunter-gatherer, agriculture and pastoral lifestyles. Evolving and adapting in Africa for the longest time compared to other global populations, Africans display the greatest genetic and phenotypic variability (Campbell and Tishkoff 2010).

The global distribution of genetic variants thus could differ from being frequently present in some populations to low or absent in others. This phenomenon could in part explain the relative differences in disease susceptibility of cancers globally (Parkin, 2004; Soerjomataram et al., 2012; Jemal et al., 2012; Chokunonga et al., 2013), and the variable incidence rates between continents (Parkin, 2004; Ferlay et al., 2010; Soerjomataram et al., 2012; Jemal et al., 2012). During his work from 1859 to 1864, August Hirsch noticed this global trend and coined the term "geographic pathology" (Beck, 1961). These aspects illustrate the need to identify genetic variability within and between populations in elucidating the role of genetic factors in diseases.

Genetic variation is present between individuals in a population and between populations. The variation is created primarily by mutations due to changes in the genetic code at a single nucleotide position (called single nucleotide polymorphism – SNP), in short tandem repeats, and in recombination where unequal transfer of genetic material can result in insertion deletion mutations (indels) of large sections of DNA code or duplication of regions. These mutations could result in altered function of genes or non-function of genes or have no effect at all. The prevalence of these mutations is the result of successful sexual transmission of the mutation from

one generation to the next. Genetic variation can be created further by influences of natural selection, population migration events and genetic drift or a combination of these factors.

The section that follows describes the histories of populations investigated in the study, with an understanding of how populations have been and are classified using linguistics, archaeology, anthropology and genetics.

1.1 The sub-Saharan African population history

Scholars have often reported the ethnicity of Africans to correlate with the language they speak, which inevitably correlates to the culture followed (Nurse et al., 1985; Cavalli-Sfroza et al., 1994; Ehret, 2000; Crawhall, 2006). Language is an expression of human culture where it is a medium to communicate with one another. In the past aspects of life events, new ideas, changes in technology and climate would have resulted in the generation of new words. In time new languages or dialects would have evolved as populations grew and ventured into different regions. By tracing these changes and identifying similarities in words used by contemporary populations, linguists have been able to trace the history of African languages. The words used would allow one to highlight aspects of your 'being', your heritage, your culture; and by this had become an identifier of your ethnicity (Ehret, 2000; Crawhall, 2006). For this reason the languages spoken by Africans cannot be treated as a discrete entities or countable units that are confined to discrete boundaries like the political structures that have been created overtime in defining countries (Ehret, 2000).

Geneticists and archaeologist have used their science to trace the histories of humans and the variation that exist among different populations. Similar to linguists, archeologists use relics of populations to understand the nature of their cultures and examine relations among populations to identify possible relations. Geneticists use DNA to trace genetic anitquities that would have been passed on from one generation to the next to understand historically the relatedness of populations to one another (Ehret, 2000; Crawhall, 2006). Cavalli-Sfroza et al. (1994) describes how by mapping the languages of the world and their genetic variation a pattern of the dispersal of Man globally can be established. Overtime populations would have ventured from a central point leaving behind the parent population and thereby creating groups speaking a different language and having acquired a slightly different genetic make-up. This theory assumes that people tend to associate more with people of their own kind, who speak the same language; and thereby would be more attracted to a mate speaking the same/similar language, following the same/similar culture and living in the a close proximity. By this populations would retain genetic traits acquired overtime since the birth of that population. This theory is consistant for most African populations.

In contrast, certain African populations have recently been influenced by the influx of populations from other regions within and outside Africa. The unions of the indigenous and the migrant groups have resulted in offspring speaking either one of the parental languages or a combination of both, and now having a combination of genetic traces to these parental populations. In Africa, such interactions have usually been unidirectional where the migrant populations have been dominant in asserting its language and culture (Nurse et al., 1985; Güldemann and Stoneking, 2008; Henn et al., 2011). By assimilation some African populations acquired new language, different technology, new methods of subsistence and culture. They have lost their

ancient cultures and languages, and some are now extinct (Nurse et al., 1985; Güldemann and Stoneking, 2008; Schlebusch, 2010a).

There are four main language groups or phyla in Africa, namely Khoisan-speakers, Niger-Congo, Nilo-Saharan and Afroasiatic (Heine and Nurse, 2000). Populations in the study speak languages that are found in the Khoisan and Niger-Congo phyla. The sections that follow describe the history of populations used in the study.

1.1.1 South Africa

1.1.1.1. Indigenous South African populations

One of the greatest known acts of human dispersal can be found in the migrations of Bantu-speaking people from the grasslands of Cameroon-Nigeria around 4,000 years ago to southern Africa about 1,500 years ago (Nurse et al., 1985). There were massive movements of farmers who followed agriculture and the pastoral lifestyles; some had the ability to fish. The movement was first across the equatorial rain forest into East Africa, and thereafter movement to the south across areas of Zambia, Zimbabwe, Tanzania, Mozambique and Botswana. People that settled in areas of South Africa, Mozambique, Zimbabwe, Zambia and Tanzania were part of the eastern stream of the "Bantu Expansion" (Guthrie, 1948; Vansina, 1984; Nurse et al., 1985; Soodyall and Jenkins, 1993; Salas et al., 2002). These migrations led to the demic diffusion and formation of several Bantu dialects. Due to the similarities in culture and language these people have been collectively called Southeastern Bantu-speakers. It must be noted that the use of 'Bantu' here is not intended to be derogatory, but to refer to groups of people who speak Bantu or Bantoid languages, a branch of the Niger-Congo language family.

In the south, Bantu-speakers came into contact with the click-speaking Khoe-San people and assimilated some of the click synonyms into their language (Nurse et al., 1985; Barnard, 1992). While the click synonyms were from Khoe-San to Bantu, population genetic studies using mitochondrial DNA and the non-recombining section of the Y-chromosome have revealed gene flow had occurred predominantly between Khoisan-speaking females and Bantu-speaking males (Soodyall et al., 2008; Schlebusch, 2010a; Naidoo et al., 2010).

Khoisan-speakers represent aboriginal populations of southern Africa, having lived in southern Africa prior to the Bantu Expansion. In the past these populations were collectively called 'Khoisan', where San groups were called 'Bushmen'. Recent discussions with these groups have highlighted their distaste at the use of Khoisan and Bushmen, where San groups prefer to be called by their self-proclaimed community names Ju|'hoansi, ǂKhomani, !Xun, etc. (Crawhall, 2006; Schlebusch, 2010b). The more appropriate terminology to use when referring to these populations is to use 'Khoisan' when referring collectively to the languages they speak; and ethnically to refer to them collectively as 'Khoe-San' (Crawhall, 2006; Güldemann and Stoneking, 2008; Schlebusch, 2010b). The latter refers to the grouping of the hunter-gatherer San populations with the pastoralist Khoe populations (Crawhall, 2006; Schlebusch, 2010b). This nomenclature is used here to refer to these populations collectively, and individually will be referred to as Khoe and San. Some language groups in the Khoisan phyla are thought to be extinct. From historic records and current data linguists have identified three major language families, namely Ju (Northern Khoe-San), Khoe (Central Khoe-San) and !Ui-Taa (Southern Khoe-San) (Güldemann et al., 2002).

1.1.1.1.1. Southeastern Bantu-speakers of South Africa

Southeastern Bantu-speakers of South Africa comprise of Xhoza, Zulu, Sotho, Tswana, Pedi and Venda speakers, predominantly collected in and around the Gauteng Province.

1.1.1.1.2. !Xun

The !Xun, previously known as Sekele (Vasekele), are another group of San who speak a dialect related to the Ju linguistic family (Crawhall, 2006; Güldemann and Stoneking, 2008). They lived in southern Angola and northern Namibia, following hunter-gatherer liefstyles. With the migration of Bantu-speakers into these regions, they learned subsistence techniques in agriculture, livestock farming and fishing from their neighbouring Bantu-speaking agricultural and pastoralist (Barnard, 1992). During the war between Angola and South-West Africa (today Namibia, then under the ruling of South Africa) from 1966-1989 many of the !Xun people were relocated to a camp called Omega Base close to the border between Namibia and Angola. After the war, many moved to South Africa in the Schmidtsdrift region. During these periods samples were collected at both locations at these different time periods, where careful attention was taken not to duplicate samples collected from individuals that had contributed to the Omega Base sample.

1.1.1.1.3. Khwe

The Khwe, previously known as the “Black Bushmen” or Kwengo (Nurse et al., 1985) were another group sampled from the Omega Camp and Schmidtsdrift, similar to that described for the !Xun. Originally they resided in southern Angola and northern Botswana. They are San who speak Kxoe, a Central Khoisan language (Güldemann et

al., 2002; Güldemann and Stoneking, 2008). They resemble their Bantu-speaking neighbours in physical appearance, though culturally and linguistically are more closely related to their San counterparts (Nurse et al., 1985). Genetically, using the paternally inherited Y-chromosome and the maternally inherited mitochondrial DNA there appears to more genetic similarity between the Khwe and southern Bantu-speakers than their San neighbours (Soodyall et al., 2008; Schlebusch, 2010a), a result similar to that found using serogenetic markers (Nurse et al., 1985). In the past these genetic similarities were thought to be due to genetic admixture between the Khwe and neighbouring Bantu-speakers. More recently though, the presence of a specific E-M35 Y-chromosome haplotype has led to conclusions that the Khwe might be descendants of an east African pastoralist group (Henn et al., 2008; Schlebusch, 2010a; Schlebuch et al., 2012). They were responsible for introducing pastoral lifestyle into present day northern Botswana. This region is infested by tsetse flies making livestock rearing impossible, and thereby it is possible they adopted the fishing and hunter-gatherer lifestyles of neighbouring San populations.

1.1.1.2. Sea-borne immigrants of South Africa

1.1.1.2.1. South African Indian

The history of Indians' arrival in South Africa occurred from the number of migrations from north-eastern, southern and western parts of India. The indentured immigrants and the Passenger Indians were two main groups of Indian immigrants that arrived in South Africa between the late 19th Century and first half of the 20th Century (Bhanaand Brain, 1990). Indentured Indians predominantly were from the north eastern and southern parts of India and spoke Hindi and Tamil or Telugu languages, respectively (Bhana and Brain, 1990). 'Passenger Indians' arrived towards the end of the indentured system. Many settled further inland in parts of South Africa, and were primarily from the western Indian origin and spoke the Gujarati

language. The South African Indian group in this study is a collection from Johannesburg and Durban, South Africa.

1.1.1.2.2. South African White

The South African White sample consists of immigrants who arrived in South Africa since the colonial periods from 1652 (Nurse et al., 1985). Many of these individuals have ancestral links to the Dutch, French, German and English populations of Europe. South African Whites today contain African derivatives of the mtDNA lineages indicative of an African maternal ancestor in their recent family history due to some admixture with Khoisan-speaking and Bantu-speaking women (Nurse et al., 1985; Schlebusch, 2010a).

1.1.1.3. South African Coloureds

The term 'Coloured' is in this study is used to identify a group of admixed South Africans. Since the arrival of Europeans to South Africa unions between predominantly European men and Khoisan-speaking or Bantu-speaking women resulted in this admixed population. Later the arrival of Malaysian slaves via Malagasay to the Cape resulted in another admixed group called the Cape Malays (Nurse et al., 1985; Schlebusch, 2010a). These groups are collectively referred to as South African Coloured. The term "Coloured" maybe considered derogatory and it should be noted that the term is used here not with that intended purpose. The Coloured group in the study were collected in the Gauteng province of South Africa and could have had a greater level of genetic input from Bantu-speaking populations than those found in the Western Cape of South Africa.

1.1.2 Botswana

1.1.2.1. Ju|'hoansi

The Ju|'hoansi live in the regions northeast of Namibia and North in Botswana. They predominantly follow hunter-gather lifestyles and speak a language derived from the Ju linguistic family, also known as Northern Khoisan languages (Nurse et al., 1985; Barnard, 1992; Güldemann et al., 2002; Crawhall, 2006; Güldemann and Stoneking, 2008). Linguists have long studied the Ju|'hoansi establishing them as one of the most ancient populations known globally (Nurse et al., 1985; Barnard, 1992; Güldemann et al., 2002; Crawhall, 2006; Güldemann and Stoneking, 2008). Further affirmation of this fact can be found in genetic studies. By studying the mtDNA, Y-chromosome and autosomal DNA of this population, geneticists have established that the Ju|'hoansi have the oldest rooted DNA variants that can be found in human DNA tree (Knight et al., 2003; Soodyall et al., 2008; Schlebusch, 2010a, Naidoo et al., 2010; Schuster et al., 2010; Henn et al., 2011).

1.1.3 Mozambique

1.1.3.1. Southeastern Bantu-speakers of Mozambique

Bantu-speakers of Mozambique share a similar history as that to South African Bantu-speakers mentioned earlier. Their settlement in Mozambique was the result of the eastern movement of the Bantu Expansion, where they too would have come into contact with the Khoe-San populations living in this region. DNA studies have shown that Mozambican Bantu-speakers share genetic relics consistent with that found in other Niger-Congo populations; and had acquired mtDNA lineages commonly found in Khoe-San women indicative of the unidirectional assimilation of Khoe-San women into their populations (Pereira et al., 2002; Salas et al., 2002). The southeastern Bantu-speaker sample of Mozambique consists mainly of people speaking Chopi, Ronga, Shangaan and Tonga dialects.

1.1.4 Namibia

1.1.4.1. Nama

The Nama were sampled from southern Namibia bordering South Africa, within the great Namakwaland area. These Khoe-speakers, also known as Northern Khoekhoe, are grouped into the larger group of Central Khoisan-speakers (Barnard, 1992; Güldemann et al., 2002; Crawhall, 2006; Güldemann and Stoneking, 2008). Historically, the Nama were nomads who followed a pastoral lifestyle. Influences introduced recently after the settlement of Bantu-speakers and Europeans into their territory have led to the disintegration of their traditional subsistence lifestyles (Barnard, 1992). Genetic data of mtDNA lineages found in the Nama indicate that they share mtDNA variants similar to that found in Bantu-speaking populations. In contrast, the paternal Y-Chromosome data indicate a closer relationship to neighbouring San populations (Knight et al., 2003; Wood et al., 2005; Schlebusch, 2010a). This is consistent with the history of this population where Bantu-speaking women had become part of Nama populations, converse to San populations where San women had been assimilated into neighbouring Bantu-speaking populations (Güldemann and Stoneking, 2008)

1.1.4.2. Dama

The Dama are a neighbouring group to the Nama and were collected in the same region as the Nama. They are an enigmatic group who physically resemble Bantu-speakers, but speak the Nama language. Genetic and serology data have showed greater similarities to Bantu-speakers than their Nama neighbours (Nurse et al., 1985; Soodyall and Jenkins, 1993; Soodyall et al., 2008; Schlebusch, 2010a).

1.1.4.3. Southwestern Bantu-speakers

During the great Bantu Expansion there was a second group who migrated west into southern Africa. There remains some debate on the route used to arrive at present day Angola and Namibia. It is postulated a few groups would have migrated along with other Bantu-speakers from East Africa to Zambia and thereafter would have split to move into the southwestern parts of Africa. Another theory is of a more direct route taken by southwestern Bantu-speakers along the west coast from the Cameroon-Nigeria regions (Guthrie, 1948; Vansina, 1984; Nurse et al., 1985; Soodyall and Jenkins, 1993; Salas et al., 2002). Southwestern Bantu-speakers in this study were collected in the north of Namibia and south of Angola and comprise of Himba, Herero and Ovambo groups.

1.1.5 Zambia

1.1.5.1. Zambian Bantu-speakers

All samples in Zambia were collected in the capital city, Lusaka. The group consists of Nyanja, Lozi, Bemba and Tonga speakers, where all language groups can be found in the Niger-Congo, Narrow Bantu-speaking, central family. Zambians Bantu-speakers like those speaking languages found in the Botatwe sub-group in the upper Zambezi Valley in southwestern parts of Zambia have been shown to contain mtDNA genetic lineages found in Khoe-San populations (Barbieri et al., 2013). The Tonga though have been shown to be genetically distant from the Botatwe speakers (Barbieri et al., 2013). Barkhan (2004) examined the mtDNAs and Y-chromosomes contained in the Lusaka sample and showed almost no DNA evidence of maternal and paternal ancestries from Khoe-San populations in this sample. This study further confirmed similarities of the maternal and paternal ancestries of the Lusaka sample to Bantu-speaking populations of west and central Africa.

1.1.6 Uganda

1.1.6.1. Ugandan Bantu-speakers

The samples from Uganda were collected from the districts of Kissoro, Kabale and Rukungiri. The group consists mainly of Bakiga, Bafumbira, Banyakole and Bahororo ethnic groups, who speak dialects classified within the Niger-Congo, Narrow Bantu-speaking, central family. Barkhan (2004) examined the mtDNAs and Y-chromosomes contained in the Ugandan sample and confirmed the maternal and paternal ancestries common to those of Bantu-speaking populations of west and central Africa.

1.1.7 Democratic Republic of Congo

1.1.7.1. Democratic Republic of Congo Bantu-speakers

The samples from the Democratic Republic of Congo (DRC) were collected from individuals residing in Luozi and Kinshasha. The group consists mainly of the Manyanga ethnic group, who speak a dialect classified within the Niger-Congo, Narrow Bantu-speaking, central family. Barkhan (2004) examined the mtDNAs and Y-chromosomes contained in the Luozi and Kinshasha sample and confirmed the maternal and paternal ancestries common to Bantu-speaking populations of west and central Africa.

1.1.8 Central African Republic

1.1.8.1. Central African Republic Pygmies

This sample of Pygmies were collected from Bangui, Central African Republic (CAR), which consists of BiAka and Babenzele (also known as Mbenzele) Pygmies who belong to the Aka subgroup (Cavalli-Sfroza, 1986). These two groups were shown to be genetically similar (Coia et al., 2004) and therefore were pooled into one group,

collectively called Pygmies. The Pygmies are historically a hunter-gatherer population who live in the equatorial rain forest of central and central-west Africa, but have since assimilated the lifestyle and language of their neighbouring farming communities. They have lost their original language and now speak Bantu languages (Grimes, 2001), yet have retained morphological features common in these individuals. The BiAka populations have been shown to contain some of the oldest DNA lineages, similar to that found in San populations, indicative of these groups being the oldest surving populations today (Coia et al., 2004; Sun et al., 2009; Patin et al., 2009). Pygmy means 'fist' in Greek and have been called such due to their short stature (Cavalli-Sfroza, 1986). The term has been considered derogatory by some and we stress our use of this name is not with this purpose, but to collectively note the biological features of a group of people who have remained (until recently) relatively isolated.

1.1.8.2. Central African Republic Ubangian-speakers

These are a group of people who were collected from the same area in the CAR like their Pygmy neighbours. They speak a Niger-Congo language called Adamawa-Ubangian, which is unrelated to the African predominant Bantu languages. Adamawa-Ubangian languages have been known to have been dispersed mainly the Northern parts of the equator extending from north-west Nigeria through Cameroon, parts of southern Chad, CAR, northern Gabon, Congo and the Democratic Republic of Congo (Vansina, 1984; Williamson and Bench, 2000). Most individuals in this sample had self-proclaimed to follow the Sangha Sangha, Gbaya and Nzakara ethnicities. These Ubangian-speakers followed mainly agricultural and pastoral lifestyles.

1.2 Genetic studies in human diseases

In the last three decades there have been considerable improvements in understanding the molecular genetics of diseases that could be attributed to single gene mutations that follows Mendelian modes of inheritance. Recent research has focused on unravelling the nature of complex human diseases (Pharoah et al., 2004; Su et al., 2013). Complex or multi-factorial diseases, such as cancer, results from the combined effects of many genetic and environmental interactions (Pharoah et al., 2004; Forsberg et al., 2013; Fransen et al., 2013). While familial, twin and linkage studies have aided in mapping candidate gene mutations that are linked to cancers, these studies have been limited in identifying risk factors that are associated with cancers. Most cancers do not follow major trends of familial inheritance patterns; consequently linkage studies have produced inconsistent findings (Parkin, 2004; Fransen et al., 2013).

To overcome these problems in predicting risk for cancer, association studies using the candidate gene approach to complex diseases have proved valuable in identifying the genetic link with large to moderate effects (Furberg and Ambrosone, 2001; Futreal et al., 2004; Forsberg et al., 2013; Su et al., 2013). Association studies make use of genotype frequency in affected individuals (cases), compared with data from unaffected individuals (controls); and where differences are significant, the locus in question could relate to varying degrees of genetic susceptibility (Futreal et al., 2004; Forsberg et al., 2013; Su et al., 2013). Identifying these biomarkers has become the focus of many association studies, in elucidating the molecular pathology of cancers (Fransen et al., 2013; Su et al., 2013). By identifying the mechanisms of cancer development and risk factors for developing cancer, it is hoped that the managment of cancer would be enhanced (Furberg and Ambrosone,

2001; Parkin, 2004; Ferlay et al., 2010; Soerjomataram et al., 2012; Jemal et al., 2012; Chokunonga et al., 2013).

The section below highlights the burden cancer presents in Africa and describes the type of cancers prevalent in Africa. Given the variability among Africans and that the underlying cause of cancer in Africans is not well known, and there is a need to investigate contributing factors more intensley.

1.3 Cancer epidemic in Africa

Historically much focus has been given to communicable diseases affecting African populations. While there is growing concern about the burden of cancer in African populations, cancer remains a low public health priority in many African countries. This is mainly due to limited resources and larger emphasis on the prevention and cure of communicable diseases such as acquired immunodeficiency syndrome (AIDS)/human immunodeficiency virus (HIV) infection, malaria, and tuberculosis (Dalal et al., 2011; Chokunonga et al., 2013). It is possible that the magnitude of the disease is not known to policy makers, the general public and other public health agencies. There are very few cancer registries throughout Africa and limited regionally, rather than nationally (Jemal et al., 2012; Chokunonga et al., 2013). To date there are three comprehensive registries which have been active in access of twenty years, namely the South African National Cancer Registry, National Health Laboratory Service; the Zimbabwean National Cancer Register and Ugandan cancer registry (Chokunonga et al., 2013). Figure 1 is an account of the cancer registry information available used in GLOBOCAN 2008, where the many red and orange highlighted countries illustrates the lack of cancer registers, with few countries having comprehensive active registries established.

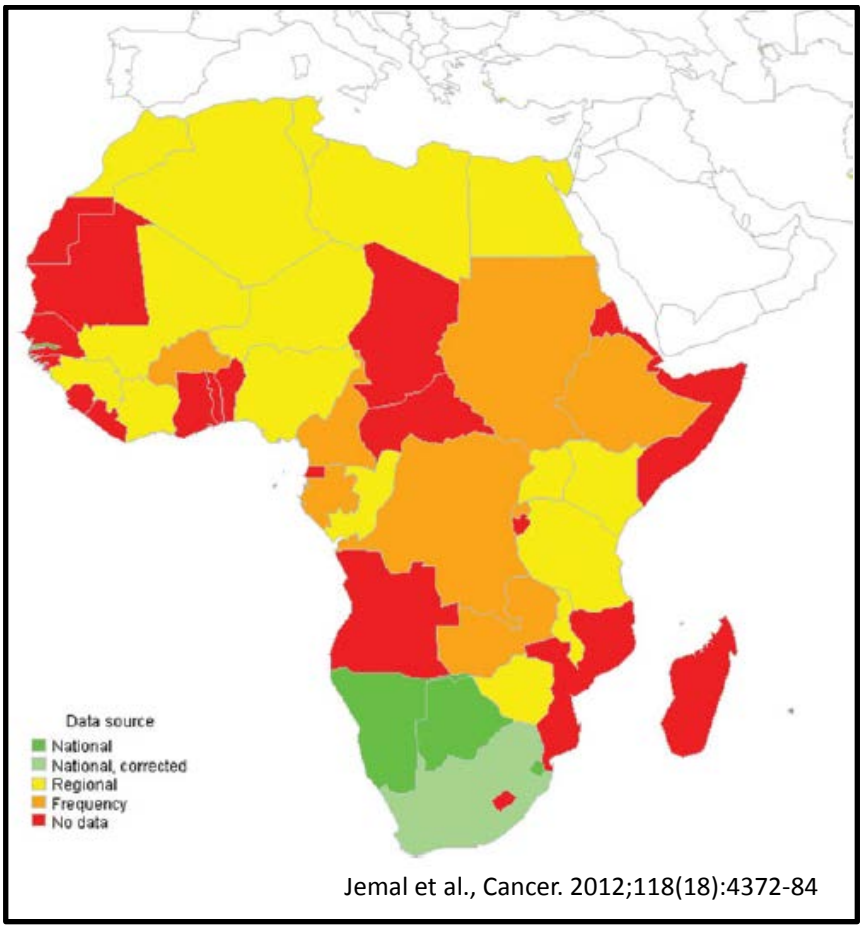

Figure 1 Methods of incidence data estimation for GLOBOCAN 2008

While GLOBOCAN 2008 had limited information from African countries (Figure 1), the study reported ~715,000 cancer cases with ~542,000 deaths (excludes non-melonoma skin cancer) in 2008 (Ferlay et al., 2010). It is important to note these statistics include between ~71,000 to ~83,000 cases presenting with Kaposi's sarcoma (Ferlay et al., 2010), where the incidence is driven by the African HIV epidemic (Dalal et al., 2011). Kaposi's sarcoma along with other common African cancers such as lung, oesophageal, oral, breast, cervical, prostate and liver cancers, is projected to increase from the ~715,000 cases observed in 2008 to ~1.3 million cases in the year 2030 (Ferlay et al., 2010; Dalal et al., 2011). The implementation of HIV prevention programs and the increased availability of anti-retroviral treatment

have reduced HIV infection rates, and it is likely that continued progress in reducing HIV incidence would result in lower incidence rates of Kaporsi's sarcoma (Chokunonga et al., 2013).

The GLOBOCAN 2008 survey further reported cancer as being the major contributor to 7.6 million deaths (Ferlay et al., 2010; Soerjomataram et al., 2012). With population growth and longer lifespans the projected incidence of cancer cases will increase from 12.7 million in 2008 to 21.4 million worldwide (Soerjomataram et al., 2012). The increase will be greatest in low to middle-income countries comprising 76% of the postulated future cases (Soerjomataram et al., 2012). Soerjomataram et al. (2012) used data collated in 2008 by GLOBOCAN (Ferlay et al., 2010) to calculate the cancer burden worldwide. The study used disability-adjusted life-years (DALYs), which is a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death. Figure 2 is an illustration of the DALYs calculated for cancer in 2008 worldwide, where the five major cancers that contribute largely to the sub-Saharan Africa DALY's are liver, breast, cervical, Non-Hodgkin lymphoma and the human immunodeficiency virus (HIV) related Karposi's sarcoma (Soerjomataram et al., 2012).

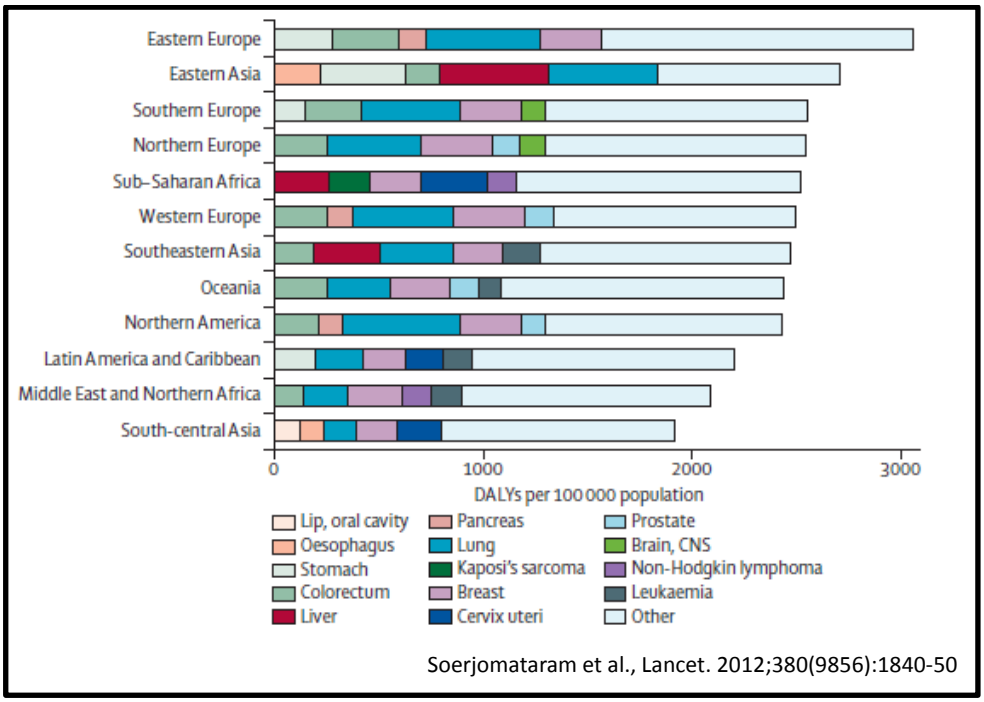

Figure 2 Age-adjusted DALY's of cancer in 2008 worldwide

Similar to global trends, the growth of the African population from 2010 to 2030 is projected to increase by 50% from 1.03 billion to 1.52 billion, so to is cancer burden projected to increase (Jemal et al., 2012; Chokunonga et al., 2013). There are several reasons for the projected increase in cancer. These include change in economic structure, relocations and change in lifestyles. Many African countries have low to middle-income economies and have growing urban lifestyles paralleled with ecomomic growth. Populations in these developing countries are relocating to urban settlements, where they are exposed to lifestyle changes such as diet, physical activity, smoking, adiposity, and alcohol use; and reproductive factors like earlier menarche, delayed childbearing and low fertility (Ferlay et al., 2010; Somdyala et al., 2010; Wentink et al., 2010; Parkin et al., 2010; Dalal et al., 2011; Jemal et al., 2012; Chokunonga et al., 2013). This shift to more Western lifestyles has been implicated in the increased burden cancer presents in African. Due to the growing concern

some African countrys' health programs have focused on determining the burden of cancer that exist in their constituencies with the aim of identifying associated risk factors (Ferlay et al., 2010; Somdyala et al., 2010; Wentink et al., 2010; Parkin et al., 2010; Dalal et al., 2011; Chokunonga et al., 2013).

Figure 3 shows the most common cancers by gender and country in Africa (Jemal et al., 2012). According to GLOBOCAN 2008 statistics breast cancer is most common (92,600 cases) among women in Africa and the second highest cause of cancer related deaths in women (50,000 deaths in 2008). Breast cancer in Africa is most common in sub-Saharan African countries, a shift from previous reports where cervical cancer was implicated. This was proposed to be due to the different reproductive factors indicated before (Ferlay et al., 2010; Jemal et al., 2012). Cervical cancer is reported to be the second most diagnosed cancer in women with 80,400 cases and 50,300 deaths. Cervical cancer rates appear to be highly prevalent in East Africa with 50 cased per 100,000 (Ferlay et al., 2010). Similarly, these trends have been reported in South Africa, where in 2006 breast cancer incidence was reported to be 29.29 per 100,000 women per year (National Cancer Registry, National Health Laboratory Service, 2006). This followed by cervical cancer at an incidence rate of 24.72 per 100,000 women per year and Basal-cell carcinoma at an incidence rate of 19.20 per 100,000 women per year (National Cancer Registry, National Health Laboratory Service, 2006).

Lung cancer is the leading cause of deaths in southern African men where it affects twice as many men as women. This has been attributed to the greater influences of tobacco consumption in southern Africa, with 65% of lung cancer cases in South Africa related to smoking (Jemal et al., 2012).

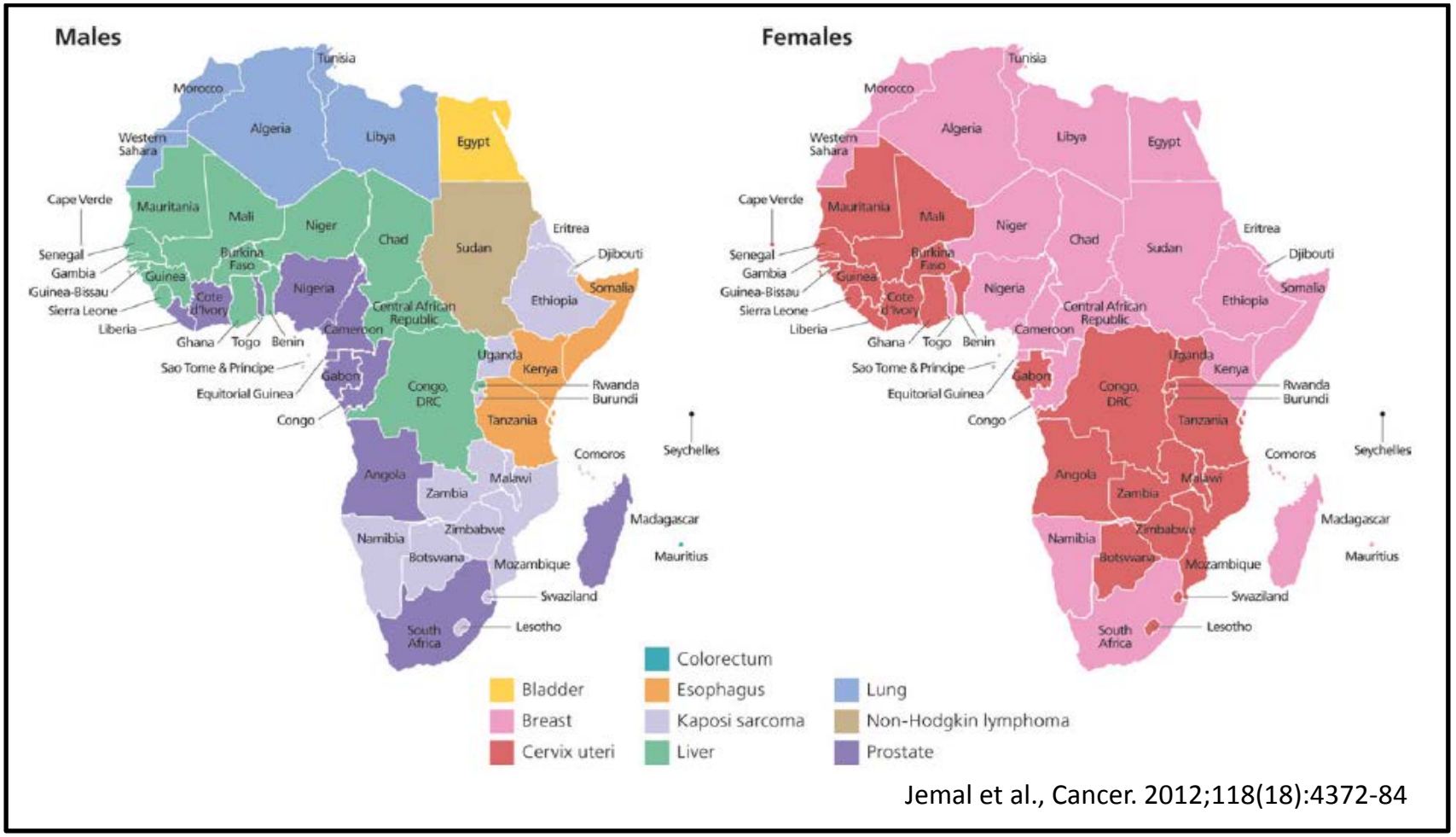

Figure 3 Most common cancer sites in Africa by sex and country are shown, GLOBOCAN 2008

Liver cancer is the second highest reported in men (34,600 cases) and third in women (16,900 cases). In 2008 33,800 men and 16,600 women died due to liver cancers. The incidence of liver cancer is compounded by the comorbidity of hepatitis B virus infections in sub-Saharan Africa and hepatitis C virus infections in North Africa, accounting for 90% of the cases (Ferlay et al., 2010).

In 2008, an estimated 37,200 cases and 30,900 deaths related to non-Hodgkin lymphoma cancers were reported (Ferlay et al., 2010). Nasopharyngeal cancer cases were estimated to be 8700 with 5500 deaths. The highest incidence rates of nasopharyngeal cancer were reported in South Africa and few North African countries (Ferlay et al., 2010). Esophageal cancer accounted for 27,900 new cases and 26,600 deaths (Ferlay et al., 2010). Squamous cell carcinomas were the predominant cases reported in the esophageal cancer bracket, largely affecting men and women of East Africa and men of South Africa. The incidence of bladder cancer was twice as high in North African men compared to southern African men, where Egyptian men having the highest incidence rates worldwide (Ferlay et al., 2010). Prostate cancer is the most common among men with 39,500 cases reported in 2008, and 28,000 deaths. Prostate cancer is most commonly diagnosed among men of southern Africa and western Africa (Ferlay et al., 2010).

1.4 Interaction of metabolic genes in cancer progression

Association studies for cancers have utilised genetic variants within metabolism genes that are involved in the detoxification of exogenous substances. Such studies within sub-Saharan African populations have been limited, where studies have either focused on one or a few mutations and/or a few populations (Dandara et al.,

2001; Dandara et al., 2002; Loktionov et al., 2002; Hamdy et al., 2003a; Dandara et al., 2003; Casson et al., 2003; Dietzsch et al., 2003; Dandara et al., 2004; Zaahl et al., 2005; Li et al., 2005; Dandara et al., 2006; Li et al., 2008; Fujihara et al., 2009; Piacentini et al., 2011). This study takes a more comprehensive look using a population genetics approach to understand the genetic variation that exists for 15 loci present in nine different genes within eighteen sub-Saharan African populations. These loci or mutations were selected as they have previously been reported to confer some genetic susceptibility to cancers globally and are primarily involved in metabolising carcinogens in two phases, namely phase I and phase II.

Phase I enzymes primarily metabolise carcinogens or other exogenous agents into different reactive states by processes of oxidation, hydrolysis or reduction. Phase II enzymes then detoxify these reactive intermediates by catalytic conjugation into water soluble products and enables excretion from the body (Furberg and Ambrosone, 2001). The interaction of these genes are crucial to the maintenance of cell integrity and the normal process of cell proliferation, where if defaulted, could result in the formation of DNA adducts and cancer progression (Futreal et al., 2004). Figure 4 describes the possible route course from carcinogen exposure to the development of cancer and the interaction of phase I and II enzymes. Further downstream DNA repair genes (not studied here) serve as another defence mechanism, where failure in function can contribute to cancer development.

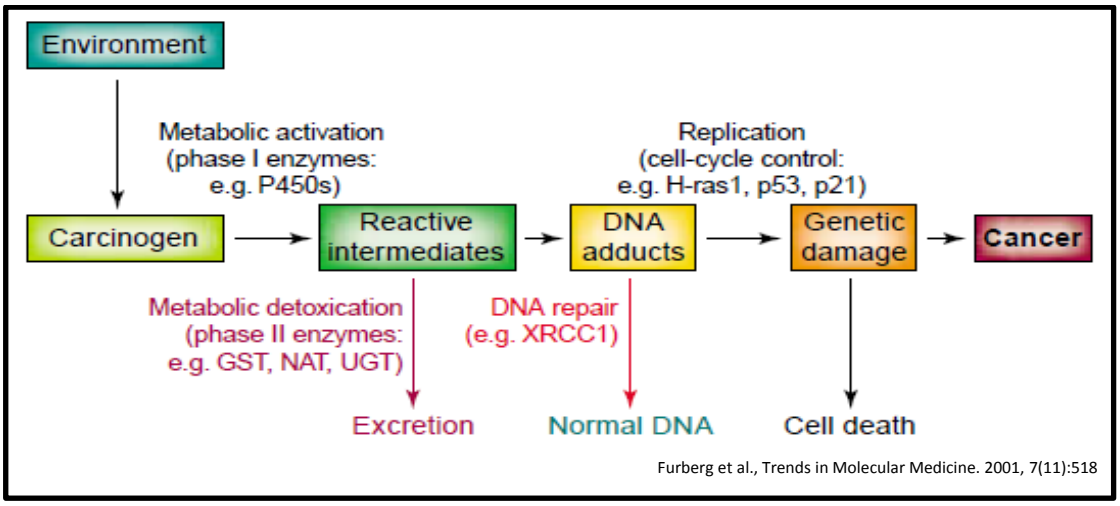

Figure 4 Pathway from carcinogen exposure to cancer development

1.5 Drug metabolising enzymes (DME)

The function of phase I and II genes and others in their superfamilies have been used in pharmacogenetic studies to understand the genetic variability that exists in global populations and partly account for the variable responses to drug treatments (Nebert, 1997; Roses, 2000; Sheweita, 2000; Peamkrasatam et al., 2006; Al-Dayel et al., 2008; Wu et al., 1998; Chaterjee et al., 2010; Brown et al., 2012; Roco et al., 2012). For many years it has been observed that patients react differently to drug treatment, where the effects of drug treatment either have no effect or are variable in response (Coutts, 1994; Nebert, 1997; Roses, 2000; Sheweita, 2000; Al-Dayel et al., 2008; Wu et al., 1998; Chaterjee et al., 2010; Roco et al., 2012). Not only are there inter-population differences, there are differences in drug metabolism between individuals. Further there is variation in drug reactivity when drugs are used in combination. The underlying reasons for this variability is relatively unknown and has become the driver to many pharmacological studies (Wu et al., 1998; Chaterjee et al., 2010; Brown et al., 2012; Roco et al., 2012). While there are several dosage, physiological and environmental factors (Roses, 2000; Sheweita, 2000), it is

postulated that drug treatment variability could be gravitated in the genetic variability in DMEs (Roses, 2000). Understanding the genetic role could lend to a greater understanding of drug metabolism and dosage curves for different populations, and enable more effective drug regiments to curb various ailments. The genes investigated in this study have various DME properties where the findings of their genetic variability could add to the knowledge of drug treatment in sub-Saharan Africa. The section that follows describes genes with DME properties that are investigated in this study, and are those previously reported to confer susceptibility to certain cancers.

1.6 Genes under investigation

For the investigation of phase I metabolisers Cytochrome P450 (CYP) superfamily genes have been selected. Genes from the Glutathione-S-transferases (GST's) superfamily and the N-acetylation 2 class gene (*NAT2*) that were selected, are grouped along in the phase II metabolising gene group.

1.6.1 Phase I genes

Cytochrome P450 (CYP) genes are a group of phase I metabolisers that play an important role in the activation of various pro-carcinogens by oxidation, which often results in the formation of highly active intermediates. Genes chosen from this superfamily are *CYP1A1, CYP1A2, CYP2D6, CYP2A6* and *CYP2E1*.

1.6.1.1 *CYP1A1* gene

CYP1A1 gene is expressed mainly in extra-hepatic tissue including the lung, where the gene codes for a well conserved enzyme that is highly activated in the metabolism of polycyclic aromatic hydrocarbons (e.g. Benzo[a]prene, main carcinogen in cigarette smoke and air pollution). For this reason, polymorphisms CYP1A1*2A (rs4646903) and CYP1A1*2C (rs1048943) have been commonly associated with increased risk to lung cancer; and oesophageal, oral and liver cancers (Wu et al., 1998; Olshan et al., 2000; Dandara et al., 2002; Ingelman-Sundberg et al., 2008; Santovito et al., 2010; Xia et al., 2013; Roco et al., 2012; Islam et al., 2013; Lopez-Cima et al., 2012; Souiden et al., 2012). There is increasing evidence for the association of these mutations with the deficient genotypes of the *GST* genes (discussed later in this chapter) and p53 tumour suppressor mutations to lung, oesophagus, and head and neck cancers (Thier et al., 2003; Xia et al., 2013). The two mutations have been widely studied, in particular many European and Asian populations (Islam et al., 2013; Lopez-Cima et al., 2012; Souiden et al., 2012).

CYP1A1*2A has been predominantly found in Japanese, Korean and Chinese populations, to lower prevalence in Caucasoid populations. In Africa, the mutation has been discovered in northern and western African populations to varying degrees where frequencies increase from North to West Africa (Okobia et al., 2005; Martinez-Labarga et al., 2007; Souiden et al., 2012). CYP1A1*2A is found at similar high frequencies in African-Americans as that found in West Africa (Olshan et al., 2000). Many studies have reported the mutation as a risk factor for lung cancer in Japanese and Caucasian populations, for colon cancer in Japanese populations, and to some degree for breast cancer in African-American populations (Wang et al., 2003; Kiruthiga et al., 2011; Xia et al., 2013; Yoshida et al., 2007; Islam et al., 2013; Souiden et al., 2012). The CYP1A1*2C mutation similarly has been found in high

frequencies in far-East Asian populations with lower frequencies in European and Indian populations. The mutation is near absent in southern African populations (Masimirembwa et al., 1998; Dandara et al., 2002) and relatively understudied throughout Africa.

1.6.1.2 *CYP1A2* gene

CYP1A2 encodes for an enzyme that activates several compounds including carcinogenic arylamines, acetaminophen, caffeine and a number of widely prescribed antipsychotic drugs such as clozapine, imipramine, caffeine, paracetamol, phenacetin, theophylline, and tacrine (Wooding et al., 2002; Fujihara et al., 2007). It has been found to be highly expressed in the liver and lesser degree in the lung (Nakajima et al., 1999; B'chir et al., 2009). More than 200 polymorphisms exists in *CYP1A2* gene, however only thirteen have been found to regulate enzyme function (Ingelman-Sundberg et al., 2008; Wang et al., 2012).

For this study two polymorphisms CYP1A2*1C (rs2069514) and CYP1A2*1F (rs762551) were chosen. These mutations have been for many years associated with lung cancer, breast cancer, colorectal cancer and stomach cancer mainly in Europeans (Nakajima et al., 1999; Sachse et al., 1999; Hamdy et al., 2003b; Yoshida et al., 2007; B'chir et al., 2009; Wang et al., 2012). However, recently a meta-analysis of such studies has not been able to corroborate these associations to cancers (Deng et al., 2013). Transcripts of CYP1A2*1C usually have low enzyme activity and for CYP1A2*1F displays increased enzyme activity (Nakajima et al., 1999). These mutations have been primarily studied in European and Asian populations. More studies have been conducted on CYP1A2*1F compared to CYP1A2*1C due to its discovery much earlier (Deng et al., 2013). CYP1A2*1C is most prevalent in Asian and

North African populations to near absence in Caucasoid populations of Europe and India (Hamdy et al., 2003b; Landi et al., 2005; Yoshida et al., 2007; Gemignani et al., 2007; B'chir et al., 2009; Singh et al., 2010). The mutation is present in Caucasian Americans and African-Americans, albeit at low frequencies (Rebbeck et al., 2007). CYP1A2*1F has been found in all populations studied globally at frequencies ranging from 40-70% in some populations to low proportions of 20-30% in some European populations (Han et al., 2001; Hamdy et al. 2003b; Cornelis et al., 2004; Soyama et al., 2005; Torres et al., 2005; Landi et al., 2005; Ketelslegers et al., 2006; Yoshida et al., 2007; Gemignani et al., 2007; Rebbeck et al., 2007; Fujihara et al., 2007; Saebø et al., 2008; B'chir et al., 2009; Sangrajrang et al., 2009; Anderson et al., 2012). The mutation is found at frequencies of 40-60% in South Africa and Zimbabwe (Dandara et al. 2004; Dandara et al., 2011).

1.6.1.3 *CYP2A6* gene

CYP2A6 is a major metabolic enzyme of nicotine. Nicotine is metabolized to cotinine by *CYP2A6*, and cotinine is further metabolized to trans-3'-hydroxycotinine, where there is a variable relationship between *CYP2A6* gene polymorphisms and inter-individual differences in nicotine metabolism (Oscarson et al., 1998; Pianezza et al., 1998; Nurfadhlina et al., 2006). For this study, polymorphisms CYP2A6*7 (rs5031016) and CYP2A6*8 (rs28399468) were chosen, where CYP2A6*7 decreases enzyme activity reducing nicotine metabolism. CYP2A6*8 effect has been noted to have no functional effect and is found to be in linkage disequilibrium with CYP2A6*7 where the combined effect decreases enzyme activity reducing nicotine metabolism (Yoshida et al., 2002; Nakajima et al., 2006; Ingelman-Sundberg et al., 2008; Djordjevic et al., 2012). The reduced nicotine metabolism diminishes nicotine dependency and in advertently could reduce risk to lung and oesophageal cancers (Pianezza et al., 1998; Yoshida et al., 2002; Nakajima et al., 2006; Han et al., 2012).

However, this had been refuted by other studies where no association between *CYP2A6* gene variants and reduced smoking was found (London et al., 1999; Chen et al., 1999; Tan et al., 2001). *CYP2A6* has further been associated with the metabolism N-nitrosodiethylamine and aflatoxin B1 (Fernandez-Salguero et al., 1995), and of pharmaceutical agents such as coumarin (+)-cis-3,5-dimethyl-2-(3-pyridl)thiazolidin-4-one hydrochloride, methoxyflurane, halothane, losigamone, letrozole, valoproic acid, disulfiram, and fadrozole (Oscarson et al., 1998; Peamkrasatam et al., 2006; Uno et al., 2013; Han et al., 2012).

Both mutations have been primarily studied in far-East Asian populations and rarely studied in other global populations (Nurfadhlina et al., 2006). Both mutations can be found in proportions less than 10% in China, Korea, and Japan, with much lower frequencies in India, Thailand and Malaysia (Nurfadhlina et al., 2006; Peamkrasatam et al., 2006; Xu et al., 2002; Djordjevic et al., 2012). From the few studies conducted in Europe, West Africa and North America, these mutations appear to be absent (Gyamfi et al., 2005; Xu et al., 2002; Djordjevic et al., 2012).

1.6.1.4 *CYP2D6* gene

The *CYP2D6* gene is involved in the detoxification of the tobacco-specific nitrosamine 4-(methylnitrosamino)-1-butanone (Krajinovic et al., 1999), and the metabolism of many drugs such as antiarhythmics, antihypertensives, 4-blockers, monoamine oxidase inhibitors, morphine derivatives, antipsychotics and tricyclic antidepressants (Linder et al., 1997; Coutts and Urichuk, 1999; Ingelman-Sundberg et al., 1999; Bradford, 2002; Lammers et al., 2010). There are about 150 polymorphisms present on the *CYP2D6* gene. However, only a few have been commonly investigated due to them rendering altered enzyme function (Bradford.,

2002). For this study CYP2D6*3 (1-bp deletion in position 2637, rs35742686) and CYP2D6*4 (G to A transition at position 1934, rs3892097) were chosen as they both have poor-metabolising properties (Sachse et al., 1997; Ingelman-Sundberg et al., 2008; Silveira et al., 2010). These genes also alter the risk to lung, bladder, gastrointestinal tract, skin cervical, breast and head and neck cancers (Krajinovic et al., 1999; Bradford, 2002; Ouerhani et al., 2008; Lammers et al., 2010; Skretkowicz et al., 2011; van der Merwe et al., 2012).

CYP2D6*3 has primarily been studied in Caucasoid populations of Europe and North America where it is found in proportions under 10% (Smith et al., 1992; Sachse et al., 1997; Leathart et al., 1998; Krajinovic et al., 1999; Gawronska-Szklarz et al., 1999; Hersberger et al., 2000; Menoyo et al., 2006; Arvanitidis et al., 2007; Buzkova et al., 2008; Correia et al., 2009; Vangsted et al., 2010; Skrętkowicz et al., 2011). The mutation is absent in Asia (Chida et al., 1999; Ji et al., 2002; Naveen et al., 2006; Nurfadhlina et al., 2006; Kiyohara et al., 2012), while evident in minute proportions in Africa and South America (Griese et al., 1999; Jorge et al., 1999; Dandara et al., 2001; Lavendera et al., 2006; Roco et al., 2012).

CYP2D6*4 has been more extensively studied globally where the mutation is found to be highly prevalent in European populations and other populations of European descent (Smith et al., 1992; Sachse et al., 1997; Leathart et al., 1998; Gonzalez et al., 1998; Krajinovic et al., 1999; Gawronska-Szklarz et al., 1999; Hersberger et al., 2000; Menoyo et al., 2006; Arvanitidis et al., 2007; Buzkova et al., 2008; Correia et al., 2009; Vangsted et al., 2010; Skrętkowicz et al., 2011; van der Merwe et al., 2012). In Asia, the mutation is absent in the far-East but present in populations of Indian descent (Chida et al., 1999; Ji et al., 2002; Naveen et al., 2006; Nurfadhlina et al.,

2006; Parveen et al., 2010; Kiyohara et al., 2012). The mutation is present at similar high frequencies in South America and the Middle East as European populations (McLellan et al., 1997; Jorge et al., 1999; Aydin-Sayitoglu et al., 2006; Kouhi et al., 2009; Gutman et al., 2009; Altayli et al., 2009; Silveira et al., 2010; Hashemi-Soteh et al., 2011; Roco et al., 2012). CYP2D6*4 is found in sub-Saharan Africa at lower frequencies than that found in other continents (Griese et al., 1999; Dandara et al., 2001; Lavendera et al., 2006; van der Merwe et al., 2012; Brown et al., 2012; Staehli et al., 2013), while higher in Tunisia as that in Europe (Ouerhani et al., 2008).

1.6.1.5 *CYP2E1* gene

The *CYP2E1* gene is a key member of the P450 suprefamily involved in the metabolism of many low molecular weight compounds and potential procarcinogens, i.e. nitrosamines, ethanol, benzene, chlorine ethylene and butadiene (Thier et al., 2003; Gordillo-Bastidas et al., 2010; Zgheib et al., 2010; Darazy et al., 2011; Huo et al., 2012). The gene is also associated with the metabolism of several pharmaceutical drugs containing acetaminophen, isoniazid, chlorzoxazone, trimethadione and d-benzphetamine, for example anaesthetics, analgesics, antipyretics, theophyllines, and chlorzoxazones (Flockhart et al., 2007; Zgheib et al., 2010; Huo et al., 2012). Due to its metabolic effects of certain procarcinogens mentioned above the gene has often been associated with a predisposition to various cancers. These cancers include nasopharyngeal, oesophageal, oral, lung, prostate and gastrointestinal tract, and various hepatic and pancreatic diseases (Lu et al., 1986; Watanabe et al., 1990; Persson et al., 1993; Mirvish et al., 1995; Hildesheim et al., 1997; Murata et al., 2001; Ferreira et al., 2003; Yang et al., 2006; Darazy et al., 2011; Huo et al., 2012). There are more than ten *CYP2E1* genetic polymorphisms that have been reported (Watanabe et al., 1990; Ingelman-Sundberg et al., 2008; Zgheib et al., 2010). For this study two polymorphic

sites of *CYP2E1* in the 5'-flanking region of the gene were chosen, *Pst*I (rs3813867) and *Rsa*I (rs2031920) in combination are classified as CYP2E1*5B. These mutations are harboured in the 5'-flanking regions of the gene have been shown to be in linkage disequilibrium. In combination these mutations increase the transcriptional activity of the gene resulting in a 10-fold increase in gene expression and increased activation of various carcinogens (Watanabe et al., 1990; Hildesheim et al., 1995; Gaspar et al., 2002; Balaji et al., 2011; Huo et al., 2012).

Both mutations have been extensively studied in East Asian populations and Europeans and farely understudied elsewhere. The mutant alleles are found to be most prominent in East Asia, and to a lesser degree in Europe and other populations of European descent (Hildesheim et al., 1995; Gonzalez et al., 1998; Rodrigo et al., 1999; Kim et al., 2000; Sarmanova et al., 2001; Kim et al., 2004; Landi et al., 2005; Chen et al., 2007; Gemignani et al., 2007; Hsieh et al., 2007; Nishino et al., 2008; Wickliffe et al., 2011; Anderson et al., 2012; Huo et al., 2012). The mutations can be found at similar high frequencies as in Europeans in India, Thailand, Russia, Turkey, Lebanon and parts of South America (Hildesheim et al., 1997; Jorge et al., 1999; Cornelis et al., 2004; Mittal et al., 2005; Aydin-Sayitoglu et al., 2006; Ulusoy et al., 2007; Gemignani et al., 2007; Da Silva et al., 2008; Zgheib et al., 2010; Coutinho et al., 2010; Sangrajrang et al., 2010; Balaji et al., 2011; Roco et al., 2012). These mutations have been relatively understudied in Africa, where data from Egypt and South Africa indicate a near absence of the mutations (Hamdy et al., 2002; Chelule et al., 2006).

1.6.2 Phase II genes

Genes of the Glutathione-S-transferases (GST's) superfamily and N-Acetylation class 2 gene (*NAT2*) are involved in the detoxification of many reactive oxidant species, often formed from the activation by phase I genes. The detoxification action forms soluble compounds that can be easily excreted from the body. Genes chosen from these phase II metabolisers are *GSTM1, GSTT1, GSTP1,* and *NAT2*.

1.6.2.1 Glutathione-S-transferase genes

Glutathione-S-transferases (GST's) are a superfamily of dimeric enzymes that detoxify a broad range of xenobiotics and carcinogens. These enzymes catalyze the reaction of glutathione with a wide variety of organic compounds to form thioethers, which are further metabolized to a more soluble form, such as mercapturic acid. The GST family is divided into four classes, namely alpha, Mu (μ), Pi (π) and Theta (θ), with each class consisting of several isoforms. Polymorphisms of three classes (μ, θ, and π) have been implicated in genetic susceptibility for head and neck squamous cell carcinomas (HNSCC), lung, bladder and many other types of cancers (Olshan et al., 2000; Casson et al., 2003; Gajecka et al., 2005; Leichsenring et al., 2006; Losi-Guembarovski et al., 2008; Torresan et al., 2008; Varela-Lema et al., 2008; Darazy et al., 2011; Lakhdar et al., 2011; Piancentini et al., 2011; Rudolph et al., 2012; Hashemi et al., 2012; Lopez-Cima et al., 2012; Liang et al., 2013).

Null alleles, GSTM1*0 and GSTT1*0, are the result of large DNA deletions in these genes that render an inability to detoxify carcinogens. Due to their detoxification properties of *GSTM1* (μ) and *GSTT1* (θ) genes, both GSTM1*0 and GSTT1*0 null alleles along with carcinogen activating cytochrome p450 genes have been implicated in cancer progression of most cancers (Olshan et al., 2000; Casson et al.,

2003; Gajecka et al., 2005; Leichsenring et al., 2006; Losi-Guembarovski et al., 2008; Torresan et al., 2008; Varela-Lema et al., 2008; Darazy et al., 2011; Lakhdar et al., 2011; Rudolph et al., 2012; Hashemi et al., 2012; Gong et al., 2012). GSTM1*0 and GSTT1*0 have been studied for their poor metabolising ability to some pharmaceutical drugs. These poor metabolising properties have been association with hepatotoxicity caused by anti-tuberculosis (TB) drugs, but there remains unconfirmed reports of this (Huang et al., 2007; Leiro et al., 2008; Lucena et al., 2008; Kim et al., 2010; Chaterjee et al., 2010; Monteiro et al., 2012).

Both null alleles have been extensively studied globally and found in all populations studied. In Africa though, many of the regions remain under investigated. GSTM1*0 appears to be more frequent (40-60%) in Europe, far East Asia, Middel East, North Africa and Caucosoid populations of North America (Olshan et al., 2000; Millikan et al., 2000; Sarmanova et al., 2001; Wang et al., 2003; Van Der Hel et al., 2005; Tsai et al., 2006; Iarmarcovai et al., 2006; Aydin-Sayitoglu et al., 2006; Yalin et al., 2007; Gundacker et al., 2007; Buchard et al., 2007; Agorastos et al., 2007; Van Emburgh et al., 2008; Liu et al., 2009; Fujihara et al., 2009; Rafiee et al., 2010; Cantor et al., 2010; Salem et al., 2011; Piacentini et al., 2011; Amer et al., 2011; Abdel-Rahman et al., 2012; Kurose et al., 2012; Rudolph et al., 2012; Lopez-Cima et al., 2012; Fischer et al., 2012). In contrast, GSTT1*0 is found at high frequencies (35-55%) in far East Asia and the few parts of sub-Saharan Africa that have been studied (Wild et al., 2000; Dandara et al., 2002; Adams et al., 2003; Wang et al., 2003; Tsai et al., 2006; Buchard et al., 2007; Liu et al., 2009; Fujihara et al., 2009; Santovito et al., 2010; Piacentini et al., 2011; Kurose et al., 2012).

GSTP1 (π) gene is a major GST expressed in the oesophagus. Two mutations, GSTP1*Ile105Val (rs1695) and GSTP1*Ala114Val (rs1138272) have been largely studied with relation to cancer predisposition. Both polymorphisms result in an amino acid change to Valine (Val), which results in reduced enzyme activity (Ali-Osman et al.1997; Mittal et al., 2005; van Emburgh et al., 2008; Matejcic et al., 2011; Gong et al., 2012). These mutations have been implicated in rendering susceptibility to head and neck cancers (including oesophageal cancer), lung, prostate and to some degree in breast cancer (Thier et al., 2003; Cho et al., 2005; Mittal et al., 2005; Li et al., 2010; Gong et al., 2012; Matejcic et al., 2011; Lopez-Cima et al., 2012). Using gene transfection and pharmacokinetics studies, *GSTP1* has been shown to be associated with the metabolism of many chemotherapeutic agents, including melphalan, cyclophosphamide, vincristine, adriamycin, cisplatin, etoposide, thiotepa, chlorambucil, and busulphan (Czerwinski et al., 1996; Awasthi et al., 1996; Hashemi et al., 2012).

GSTP1*Ile105Val is found globally at varying proportions, where it is most frequent (10 to 50%) in Africa (Masimirembwa et al., 1998; Dandara et al., 2002; Adams et al., 2003; Yen-Revollo et al., 2009; Lakhdar et al., 2010; Li et al., 2010; Matejcic et al., 2011), to having having lower frequencies (<30%) in Europe, Asia, the Americas and Middle East (Watson et al., 1998; Olshan et al., 2000; Sarmanova et al., 2001; Ballerini et al., 2003; Wang et al., 2003; Landi et al., 2005; Ketelslegers et al., 2006; Gemignani et al., 2007; Schwartzbaum et al., 2007; Altayli et al., 2009; Coutinho et al., 2010; Galvan et al., 2011; Su et al., 2012; Wang et al., 2011; Kiyohara et al., 2012; Hashemi et al., 2012; Vural et al., 2012). African Americans contain this polymorphism in almost similar proportions to their African relatives (Olshan et al. 2000).

The GSTP1*Ala114Val has been studied to a lesser degree; where the mutation is present in Africa to frequencies between 10-20% (Lakhdar et al., 2010; Li et al., 2010), and <10% in other populations globally (Watson et al., 1998; Ballerini et al., 2003; Landi et al., 2005; Ketelslegers et al., 2006; Gemignani et al., 2007; Schwartzbaum et al., 2007; Wang et al., 2011; Vural et al., 2012).

1.6.2.2 N-acetylation genes

N-acetylation (NAT) genes, *NAT1* and *NAT2* are another group of phase II metabolizers. For the purpose of this study the *NAT2* gene was selected, where it is responsible for the N-acetylation of certain aryl amines, xenbiotics and tobacco-related carcinogens (Hein, 2002; Touré et al., 2012). There appears to be bimomodal distribution of the NAT2 phenotypes globally, where individuals either have rapid acetylator phenotypes, or slow acetylator phenotypes with reduced or deficient enzyme activity (Cavaco et al., 2003; Touré et al., 2012). While *NAT2* gene has over 13 single nucleotide polymorphisms, the NAT2*14A polymorphism was selected for this study as it is found almost exclusively in African populations (Delomenie et al., 1996; Bayoumi et al., 1997; Butcher et al., 2002; Adams et al., 2003; Cavaco et al., 2003; Yoshida et al., 2007; Sabbagh et al., 2008; Dandara et al., 2010; Staehli et al., 2013; Touré et al., 2012), and absent in other global populations not of African descent (Lin et al., 1994; Cascorbi et al., 1999; Gross et al., 1999; Hein, 2002; Jorge-Nebert et al., 2002; Landi et al., 2005; Al-Moundhri et al., 2007; Staehli et al., 2013). Its localization to Africa is further evident by its presence in African Americans and Brazilians who both share an African heritage (Lin et al., 1994; Teixeira et al., 2007).

NAT2*14A renders poor metabolising function where it has been associated with the development of urinary bladder, colorectal, breast, head and neck, lung and

prostate cancers (Hein, 2000; Butcher et al., 2002). The pharmacogenetic properties of *NAT2* are in the N-acetylation of Isoniazids (INH), one of the most important drugs used in anti-TB treatment and in TB chemoprophylaxis (Teixeira et al., 2007). *NAT2* variation exists between global populations, where mutations of *NAT2* could render rapid acetylation, resulting in low concentrations of INH and possible treatment failures.

1.7 Study objectives

Given the large genetic variability present within Africa and the lack of studies in Africa, understanding the evolutionary histories of these populations and the genetic variability of genes coding xenobiotic metabolising enzymes would address fundamental questions within the scientific literature. Such research would aid in our understanding of genetic factors associated with multifactorial diseases, and could give a better understanding of the associated pharmacological properties. Acknowledging the evolution and demographic history that has shaped African populations, this study was designed to use a population genetics approach to:

(i) Examine the allelic variation within the *CYP1A1, CYP1A2, CYP2A6, CYP2D6, CYP2E1, GSTM1, GSTT1, GSTP1* and *NAT2* genes in some sub-Saharan African populations;

(ii) Use the frequencies and distribution of variants of these candidate genes to examine the genetic affinities within 18 sub-Saharan African populations;

(iii) Evaluate the patterns of variation within Africa to non-African populations;

(iv) Assess relevance of candidate loci for future studies on cancer and pharmacogenomics in Africa.

CHAPTER 2 SUBJECTS, MATERIALS AND METHODS

2.1 Subjects

The DNA samples from a total of 1880 subjects (males and females) residing in eight African countries: South African, Botswana, Mozambique, Namibia, Zambia, Uganda, Democratic Republic of Congo (D.R.C.), and Central African Republic (C.A.R.). Figure 5 illustrates the locations where samples were collected. Eighteen population groups were classified by their geographic location, ethnicity and linguistic affiliation (Table 1). A description of these population histories can be found in section 1.1.

The samples were collected for population genetic studies in sub-Saharan Africa by Professor Soodyall and colleagues in the Division of Human Genetics (National Health Laboratory Services and University of the Witwatersrand) (Protocol number M050906, renewed 29 May 2009, Appendix A). The use of these samples for this study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Protocol number M060325; Appendix A).

All samples were collected after seeking informed consent from volunteers. At the time of sampling volunteers were asked to provide information relating to their place of birth, home language and that of parents (and in some cases grand-parents), and their self-proclaimed ethnic identity. Having reviewed the various factors when defining African populations (see Section 1.1), it was decided to rather group the sample according to individuals self-proclaimed ethnic identity, mapped to the countries in which they had been collected (Figure 5 and Table 1).

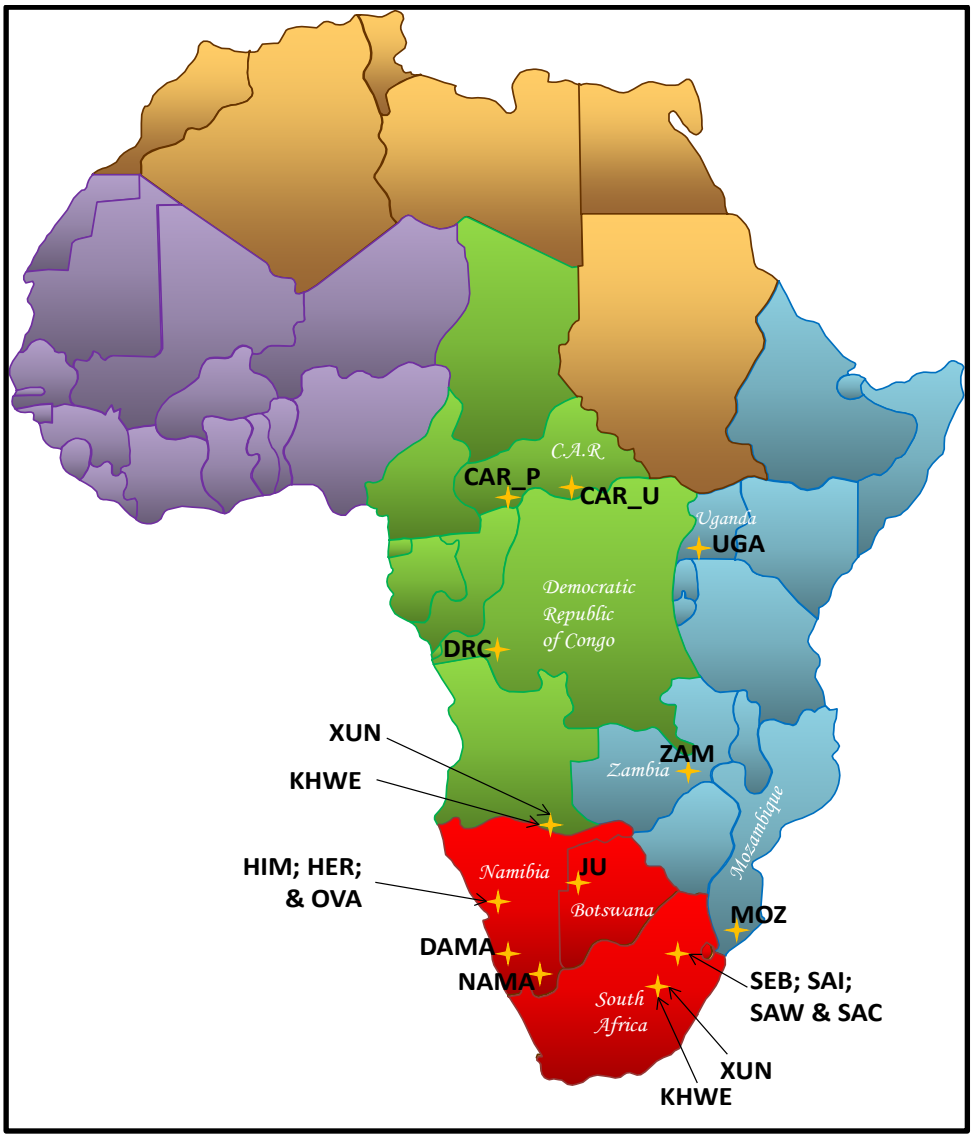

Figure 5 Map indicating position where samples were collected

Population code key: XUN-!Xun; KHWE-Khwe; SEB-South African Bantu-speakers; SAI-South African Indians; SAW-South African Whites; SAC-South African Coloureds; MOZ-Mozambican Bantu-speakers; NAMA-Nama; DAMA-Dama; JU- Ju|'hoansi; HIM-Himba; HER-Herero; OVA-Ovambo; ZAM-Zambian Bantu-speakers; DRC-Democratic Republic of Congo Bantu-speakers; UGA-Ugandan Bantu-speakers; CAR_U-Central African Republic Ubangian-speakers; CAR_P-Central African Republic Pygmies.

XUN and KHWE populations were sampled at different time periods and at two different locations. Colour highlights of countries represent the different regions of Africa: red – Southern Africa; blue – Eastern Africa; green – Central Africa; purple – Western Africa; and brown – Northern Africa

Table 1 Study population-sample origin, ethnicity, language, population code and sample size

Sample Origin by Country	Population Name	Ethnicity(ies) €	Language Classification	Population Group Code§	N
South Africa	South African Bantu-speakers ¥	Southeastern Bantu-speakers (Xhoza, Zulu, Sotho, Tswana, Pedi and Venda)	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	SEB	115
South Africa & Namibia#	!Xun ¥	San	Northern Khoisan: !Xun [2]	XUN	76
South Africa & Namibia#	Khwe ¥	San	Central Khoisan: Khwe [2]	KHWE	49
South Africa	South African Indians ¤	Tamil, Telugu, Hindi and Gujarati	Indo-Arian & Dravidic	SAI	114
South Africa	South African Whites ¤	Anglo, Dutch, French and Germanic	English & Afrikaans	SAW	51
South Africa	South African Coloureds	Mixed	English & Afrikaans	SAC	91
Botswana	Ju|'hoansi ¥	San	Northern Khoisan: Ju|'hoansi [2]	JU	30
Mozambique	Mozambican Bantu-speakers ¥	Eastern Bantu-speakers (Chopi, Ronga, Shangaan and Tonga)	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	MOZ	70
Namibia	Nama ¥	Khoe	Central Khoisan: Nama [2]	NAMA	22
Namibia	Dama** ¥		Central Khoisan: Nama [2]	DAMA	37
Namibia	Himba ¥	Southwestern Bantu-speakers	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	HIM	40
Namibia	Herero ¥	Southwestern Bantu-speakers	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	HER	79
Namibia	Ovambo ¥	Southwestern Bantu-speakers	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	OVA	62
Zambia	Zambian Bantu-speakers ¥	Eastern Bantu-speakers (Nyanja, Lozi, Bemba and Tonga)	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	ZAM	179
Uganda	Ugandan Bantu-speakers ¥	Eastern Bantu-speakers (Bakiga, Bafumbira, Banyakole and Bahororo)	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	UGA	245
Democratic Republic of Congo	D.R.C Bantu-speakers ¥	Central Bantu-speakers (Manyanga/Nyanga)	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central[1]	DRC	302
Central African Republic	C.A.R. Pygmy ¥	Central Bantu-speakers (BiAka and Babenzele)	Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Northwest[1]	CAR_P	120
Central African Republic	C.A.R. Ubangian-speakers ¥	Central Bantu-speakers (Sangha Sangha, Gbaya and Nzakara)	Niger-Congo, Atlantic-Congo, Volta-Congo, North, Adamawa-Ubangi, Ubangi[1]	CAR_U	198

1 http://www.ethnologue.com/country/ZA/languages

2 Güldemann et al., 2000; Güldemann et al., 2008

€ Most African ethnicities correlate with language dialects spoken

These samples were collected at two locations: In Namibia (then South-West Africa) at the Omega Camp and in South Africa in Schmidtsdrift region

**Enigmatic group who differ from the Nama in physical appearance but have adopted the Nama language

§ These codes will be used throughout in refence to the populations sampled

¥ Indigenous African population

¤ Sea-borne immigrants population

2.2 Comparative data

Where available, data from other populations were used to compare frequencies of alleles at the genetic loci used in the current study. African populations have been divided into different regions as indicated in Figure 5. Comparative data collated for this study along with references to the studies used can be found in tables in Appendix B.

2.3 Materials and methods

2.3.1 DNA extractions and quantification

Where blood samples were collected, the DNA was extracted using the salting out procedure described by Miller et al. (1988). Where buccal swabs were collected, DNA was extracted according to manufacturer's specifications using commercial available kit manufactured by Gentra Systems called the PureGene® Genomic DNA Purification.

DNA samples were quantified spectrophotometrically using the NanoDrop ND-1000 Spectrophotometer (ColemanTechnologies Inc., LabVIEW®). Quantified DNA were then diluted to the required concentration and volume by diluting the DNA sample in sterile double distilled water (ddH_2O).

2.3.2 Null allele typing methods

To determine the presence of null mutations GSTM1*0 and GSTT1*0 in the sample, the method reported by Casson et al. (2003) was used and optimised. This is a

multiplex polymerase chain reaction (PCR) where primers (Table 2) amplify across these deleted regions for the *GSTM1* and *GSTT1* genes. In addition, primers for a control band in the β-Globin gene was included as a positive control in the event an individual contained both GSTM1 and GSTT1 null alleles - to discriminate between a failed PCR and both mutations being absent. The primer sequences used in this multiplex reaction to detect null mutations are given in table 2.

Table 2 Primer name and sequences for GSTM1 and GSTT1 null allele detection

Target region	Primer name	Primer sequence (5' to 3')
GSTM1 null allele	ForM1	CTTGGGCTCAAATATACGGTGG
	RevM1	GAACTCCCTGAAAAGCTAAAGC
GSTT1 null allele	ForT1	TCACCGGATCATGGCCAGCA
	RevT1	TTCCTTACTGGTCCTCACATCTC
β-Globin control	β-GlbFor	GAAGAGCCAAGGACAGGTAC
	β-GlbRev	AACTTCATCCACGTTCACC

About 100ng of DNA of each sample was used in the multiplex PCR. In addition to DNA, each PCR mix contained 0.2µM of each primer (Table 2), 2.5µl FastStart Taq Buffer solution containing 1.5mM $MgCl_2$, 2U of FastStart Taq (Roche Applied Science), and 10µM dNTP mix (containing dATP, dGTP, dCTP, dTTP), all made up to a final volume of 25µl with ddH_2O. The thermal cycler program was as follows: one cycle at 95˚C for 5 minutes; followed by 30 cycles at 95˚C for 30 seconds, 61˚C for 30 seconds, and extension at 72˚C for 1 minute; with a final extension step at 72˚C for 7 minutes. Thereafter, 10µl of the PCR product was resolved on a 2% Agarose gel

stained with ethidium bromide (electrophoresis conditions: 1 x TBE running buffer, a loading dye, 1Kb DNA ladder [Gibco BRL] size standard).

2.3.3 Detection of SNP markers using the Single Base Extension method

The Single Base Extension (SBE) method is a minisequencing procedure which involves:

I. Target regions containing the SNPs are first amplified using PCR either in single reactions or multiplex reactions. PCR products are then cleaned to remove all unincorporated PCR primer and other PCR components, retaining amlpicons only.

II. These purified PCR products are then used in a minisequencing reaction along with probes (sense or anti-sense primer) designed to anneal adjacent to the target SNP and fluorescently tagged ddNTPs (whereby each ddNTP is labelled with one of four fluorescent coloured dyes). The different probes are in addition designed to vary in length by adding $(GATC)_n$ primer tails, where by size fractionation during electrophoresis it is possible to distinguish the different SNP loci being typed. The reaction follows where PCR products are denatured to become single strands to allow for probes to anneal to regions adjacent to SNP. In an elongation step the corresponding ddNTP to the target SNP binds to the probe. After several cycles probes would now contain ddNTPs that correspond to the SNP targeted. The products of this reaction are then cleaned to remove any unincorporated components to retain probes that contain ddNTP.

III. These minisequencing reaction products can now be resolved using electrophoresis. A capillary array DNA analyser is often used to resolve probe fragments, where fluorescent excitation enables the detection of the ddNTP insert. Depending on the colour emmited by one of four fluorescent dyes and

the size fractionation of the peaks, the allele present for a particular SNP can be determined.

In this study the SBE method was used to detect 13 SNP markers implicated in genetic susceptibility to cancer, where simultaneously an internal control band from the β-Globin gene was included. Table 3 lists the SNP markers used, the corresponding multiplex PCR primers with sequences, the SBE primer (probe) sequences and primer sizes with poly(GATC)$_n$ extensions, PCR product sizes, and the allele discrimination identifiers.

2.3.3.1 Primer design

The target regions containing the relevant SNPs were derived from DNA sequences of the GenBank database (refer to Appendix C). The Primer3 software (Rozen and Skaletsky, 2000) tool was used to design PCR and SBE primers. Primers were then aligned to the NCBI BLAST alignment tool to confirm template specificity. To rule-out primer-dimers and hairpin loop formation AutoDimer was used (Vallone and Butler, 2004). Primers were ordered from Metabion. To ensure high quality and specificity in DNA amplification HPLC-purification was requested for each primer.

PCR primer sets were designed to anneal at temperatures ranging from 56°C to 61°C to allow for some flexibility to aid in the optimization of amplifying all fragments optimally in a single multiplex PCR. PCR primer lengths were between 18 and 23 mers (Table 3); where the GC content of primers varied from 39.1% to 61.9%. Primers were designed to amplify regions varying from 124bp to 340bp (Table 3), where the varying lengths would aid in differentiating products after being resolved by agarose gel electrophoresis.

SBE primers were designed to anneal at low temperatures (~50°C) with high specificity to the binding region alongside the SNP; with varying lengths by adding poly-GATC tails of differing lengths to the 5' end of all SBE primers, except one (Table 3). The SBE primers were carefully designed to allow adequate differentiation during capillary electrophoresis, where lengths varied between 26 to 117 mers. The design of SBE primer lengths is important where two or more fragments if they migrate together through the capillary could result in an output where dye peaks could lie on top of one-another and make the result undecipherable, especially if the fluorescent dye emission is of the same colour.

2.3.3.2 Multiplex PCR optimization

To amplify the 13 PCR fragments and the internal control band, fragments were first amplified in separate reactions at varying annealing temperatures ranging between 54°C to 60°C using a gradient thermocycler. Products for each annealing temperature variant were then resolved on a 3% agarose gel stained with ethidium bromide. These initial reactions were done to identify the most optimal conditions for the each primer pair, and thereby identifying an optimal condition at which all primers pairs could be amplified in a multiplex reaction.

For each primer pair tested (Table 3), each PCR contained about 10ng of DNA, 0.2µM of each primer (Table 3), 2.5µl FastStartTaq Buffer, 1.5mM $MgCl_2$, 1U of FastStartTaq (Roche Applied Science), and 15µM dNTP mix (containing dATP, dGTP, dCTP, dTTP), all made up to a final volume of 25µl with ddH_2O. The thermal cycler program was as follows: one cycle at 95˚C for 5 minutes; followed by 30 cycles at 95˚C for 30 seconds, primer annealing temperature range between 54°C and 60°C

(as set by temperature gradient of the gradient thermocycler) for 30 seconds, and extension at 72˚C for 1 minute; with a final extension step at 72˚C for 7 minutes.

Once optimal conditions were determined individually, all primer sets were combined in a multiplex reaction. The multiplex reaction was further optimized by varying primer set concentrations, the FastStart Taq Buffer concentration, $MgCl_2$ concentrations, using annealing temperatures between 54°C and 57°C, and DNA template amounts of 8ng, 10ng and 25ng. The final optimized reaction contained about 10ng of DNA, 1µl of primer mix (see table 4 for primer concentrations), 2.5µl FastStart Taq Buffer, 2.0mM $MgCl_2$, 1U of FastStart Taq (Roche Applied Science), and 30µM dNTP mix (containing dATP, dGTP, dCTP, dTTP), all made up to a final volume of 25µl with ddH_2O. The thermal cycler program was as follows: one cycle at 95˚C for 5 minutes; followed by 35 cycles at 95˚C for 30 seconds, 57°C for 30 seconds, and extension at 72˚C for 1 minute; with a final extension step at 72˚C for 7 minutes.

PCR products were hereafter purified by adding 1.4U of Shrimp Alkaline Phosphatase (USB Corporation) and 4U of Exonuclease I (New England Biolabs) to 5µl PCR product made up to a total reaction volume of 7µl with ddH_2O. The reaction was incubated at 37˚C for 60 minutes followed by an enzyme inactivation step of 75˚C for 15 minutes.

Table 3 List of SNP markers typed using SBE reaction along with primer sequences, fragment lengths, and alleles colour coded to the respective ddNTP fluorescent dye contained in the SBE reaction mix

Locus	Gene	Primer Direction	Primer Sequence ('5 - 3')	Final concentration in PCR (μM)[#]	PCR product size (bp)	SBE primer name	$(gatc)_n$ + SBE primer sequence (5' - 3')	Final concentration in SBE reaction (μM)[β]	Total SBE primer length (bp)	[§]Alleles: colour coded by floresent dye colour
Control band	B-Globin	Forward	GAAGAGCCAAGGACAGGTAC	0.10	267	β-Glb_SBE	$(gatc)_7$tggtgcatctgactcctgag	0.07	48	G
		Reverse	AACTTCATCCACGTTCACC	0.10						
GSTP1*Ile105Val	GSTP1	Forward	ACCCCAGGGCTCTATGGGAA	0.10	176	GSTP105_SBE	$(gatc)_{10}$aggacctccgctgcaaatac	0.20	60	A allele (wild type)
		Reverse	TGAGGGCACAAGAAGCCCCT	0.10						G allele (mutant)
GSTP1*Ala114Val	GSTP1	Forward	GTTGTGGGGAGCAAGCAGAGG	0.05	216	GSTP114_SBE	$(gatc)_{20}$cttcacatagtcatccttgccc	0.06	102	G (C) allele (wild type)
		Reverse	CACAATGAAGGTCTTGCCTCCC	0.05			anti-sense primer[§]			A (T) allele (mutant)
CYP1A1*2A	CYP1A1	Forward	CAGTGAAGAGGTGTAGCCGC	0.07	340	CYP1A1*2A_SBE	$(gatc)_{10}$tgtttcactgtaacctccacctcc	0.20	64	T allele (wild type)
		Reverse	TAGGAGTCTTGTCTCATGCC	0.07						C allele (mutant)
CYP1A1*2A	CYP1A1	Forward	CAGCTGTCTCCCTCTGGTTA	0.04	185	CYP1A1*2C_SBE	$(gatc)_{15}$ggaagtgtatcggtgagacc	0.06	80	A allele (wild type)
		Reverse	GCCAGGAAGAGAAAGACCTC	0.04						G allele (mutant)
CYP1A2*1C	CYP1A2	Forward	GCTACACATGATCGAGCTATAC	0.05	155	CYP1A2*1C_SBE	$(gatc)_5$ccgcaacctccgcctctc	0.07	38	G allele (wild type)
		Reverse	GCATGACAATTGCTTGAATC	0.05						A allele (mutant)
CYP1A2*1F	CYP1A2	Forward	GATGATGTGTGGAGGAGAGA	0.05	225	CYP1A2*1F_SBE	$(gatc)_{18}$aaagggtgagctctgtgggc	0.12	92	C allele (wild type)
		Reverse	AGGGTTGAGATGGAGACATT	0.05						A allele (mutant)
CYP2A6*7[Ψ]	CYP2A6	Forward	CTCAAGTCCTCCCAGTCACC	0.10	241	CYP2A6*7_SBE	$(gatc)_{24}$acgtgtttgggggacacgtca	0.12	117	A (T) allele (wild type)
							anti-sense primer[§]			G (C) allele (mutant)
CYP2A6*8[Ψ]	CYP2A6	Reverse	TCTCTTCCCTCTAGCCACCA	0.10		CYP2A6*8_SBE	$(gatc)_8$ggcaggaagctcatggtgtagttt	0.16	56	C (G) allele (wild type)
							anti-sense primer[§]			A (T) allele (mutant)

Table 3 continued

Locus	Gene	Primer Direction	Primer Sequence ('5 - 3')	Final concentration in PCR (μM)[#]	PCR product size (bp)	SBE primer name	$(gatc)_n$ + SBE primer sequence (5' - 3')	Final concentration in SBE reaction (μM)[β]	Total SBE primer length (bp)	[§]Alleles: colour coded by floresent dye colour
CYP2D6*3A (2549delA)	CYP2D6	Forward	TAGGTGCTGAATGCTGTCC	0.05	185	CYD2D6*3A_SBE	$(gatc)_{22}$gatgagctgctaactgagcac	0.16	109	A allele (wild type)
		Reverse	CTCACCTTCTCCATCTCTGC	0.05						G allele (mutant)
CYP2D6*4 (1846G>A)	CYP2D6	Forward	GTGGGTGATGGGCAGAAG	0.08	124	CYD2D6*4_SBE	$(gatc)_2$ccgcatctcccacccca	0.20	26	G allele (wild type)
		Reverse	CTTTGTCCAAGAGACCGTTG	0.08						A allele (mutant)
NAT2*14A	NAT2	Forward	TTAGGGGATCATGGACATTG	0.08	227	NAT2*14A_SBE	$(gatc)_{12}$tgatcacattgtaagaagaaacc	0.07	71	G allele (wild type)
		Reverse	GATTGACCTGGAGACACCAC	0.08						A allele (mutant)
CYP2E1*5B(*PstI*)	CYP2E1	Forward	CACTGGAAAGGAAAGAGAGG	0.12	206	CYP2E1*PstI_SBE	$(gatc)_6$cccttcttggttcaggagag	0.20	44	G allele (wild type)
		Reverse	GCCACATAAGCAAGTCATTG	0.12						C allele (mutant)
CYP2E1*5B(*RsaI*)	CYP2E1	Forward	AAGTGATTTGGCTGGATTGT	0.05	237	CYP2E1*RsaI_SBE	$(gatc)_{15}$gttcttaattcataggttgcaatttt	0.15	86	G (C) allele (wild type)
		Reverse	CCCTCTTCCACCTTCTATGA	0.05			anti-sense primer[§]			A (T) allele (mutant)

§ Anti-sense sequence used to design SBE primer, with the corresponding allele is highlighted in purple print in the last column

¥ Is the only SBE primer without a poly(gatc) tail

Ψ One PCR amplicon containing SNPs for two SBE primers

Primer concentration in a 25μl PCR volume

β SBE primer concentration in a 5μl SBE reaction volume

2.3.3.3 Genotyping of SNP markers using the SBE method

Following multiplex PCR amplification and purification, 1.5µl of the purified PCR product was used in a SBE reaction. The SBE reaction mix consisted of 1µl of ABI PRISM® SNaPshot™ Multiplex Ready Reaction Mix, 1µl of the SBE primer cocktail (see Table 4), and made up to a volume of 5µl with ddH_2O. With each SBE amplification round, a positive control reaction (amplified separately) supplied with the SNaPshot™ Multiplex Ready Reaction Mix was included. Amplification conditions were 35 cycles of a denaturation step at 96˚C for 10 seconds, SBE primer annealing at 50°C for 5 seconds and the single base ddNTP elongation step at 60˚C for 30 seconds.

Following the SBE reaction, products were purified to retain SBE probes with fluorescently tagged ddNTP incorporated by adding 0.5U of Shrimp Alkaline Phosphatase (USB Corporation) to 5µl of the SBE reaction products, all made up to a total reaction volume of 7µl with ddH_2O. The reaction was incubated at 37˚C for 60min, followed by 75˚C for 15 minutes.

To genotype SBE reaction samples, 2µl of cleaned SBE products were mixed with 7.5µl Hi-Di formamide (Applied Biosystems) and 0.5µl of GeneScan-LIZ 120 internal size standard (Applied Biosystems). Each prepared mixture was then denatured briefly at 95˚C for 2 minutes and quickly cooled to 4˚C. Similar mix and treatment was done for the positive control supplied with the ABI PRISM® SNaPshot™ Multiplex Ready Reaction. SBE products were then resolved using ABI PRISM® 3130xl Genetic Analyzer (Applied Biosystems) according to ABI PRISM® SNaPshotTM Multiplex Kit instructions (Applied Biosystems) and analyzed using GeneMapperID v3.2 software (Applied Biosystems).

Table 4 Multiplex PCR primer and SBE primer concentrations used in SBE method

Primer name	Direction	Final concentration (μM)[¥]	SBE Primer name	Final concentration (μM)[Ψ]
β-Glb-F	Forward	0.10	β-Glb_SBE	0.07
β-Glb-R	Reverse	0.10		
P105F	Forward	0.10	GSTP105_SBE	0.20
P105R	Reverse	0.10		
P3	Forward	0.05	GSTP114_SBE	0.06
P4	Reverse	0.05		
CYP1A1*2F	Forward	0.07	CYP1A1*2_SBE	0.20
CYP1A1*2R	Reverse	0.07		
CYP1A1*3F	Forward	0.04	CYP1A1*3_SBE	0.06
CYP1A1*3R	Reverse	0.04		
CYP1A2-R2	Forward	0.05	CYP1A2-1C_SBE	0.07
CYP1A2-Cr	Reverse	0.05		
CYP1A2-Ff	Forward	0.05	CYP1A2-1F_SBE	0.12
CYP1A2-Fr	Reverse	0.05		
CYP2A6*7F	Forward	0.10	CYP2A6*7_SBE[§]	0.12
CYP2A6*7R	Reverse	0.10	CYP2A6*8_SBE[§]	0.16
CYP2D6*3F	Forward	0.05	CYP2D6*3_SBE	0.16
CYP2D6*3R	Reverse	0.05		
CYP2D6*4F	Forward	0.08	CYP2D6*4_SBE	0.20
CYP2D6*4F	Reverse	0.08		
NAT2*14F	Forward	0.08	NAT2*14_SBE	0.07
NAT2*14R	Reverse	0.08		
NAT1*10F	Forward	0.20	NAT1*10_SBE	0.16
NAT1*10R	Reverse	0.20		
CYP2E1-PsF	Forward	0.05	CYP2E1-PstI_SBE	0.15
CYP2E1-PsR	Reverse	0.05		
CYP2E1-RsF	Forward	0.12	CYP2E1-RsaI_SBE	0.20
CYP2E1-RsR	Reverse	0.12		

§ One PCR amplicon containing SNPs for two SBE primers
¥ Primer concentration in a 25μl PCR volume
Ψ SBE primer concentration in a 5μl SBE reaction volume

A macro was designed in the GeneMapperID v3.2 software (Applied Biosystems) according to manufacturer's recommendations where the genotypes of each SBE fragment would be automatically scored and an allele output per marker would be generated. In table 3, the last column lists the alleles for each SNP and is colour coded to represent the respective fluorescent dye colour that would be presented on electropherogram outputs. Table 3 in addition displays the expected fragment size of the SBE product. The macro designed uses the GeneScan-LIZ 120 internal size standard (Applied Biosystems) to determine the SBE product size relative to the migration in the electrophoresis run, where the size of the fragment serves as an identifier of a SNP being detected.

2.4 Statistical analysis

Genotypes for the null allele PCR were read off the gel images taken after PCR products were resolved on agarose gels. The presence or absence of the null alleles in the *GSTM1* and *GSTT1* genes were entered into EXCEL spreadsheets. From these collated data tables genotype frequency calculations were computed for each population.

SNP markers were typed using the SBE technique described here. The GeneMapperID v3.2 software was used to analyse the outputs of these genotyping runs, where the samples were automatically scored to the respective genotype for the respective region, and hereafter tabulated in EXCEL spreadsheet. The tabulated results were checked against the electropherograms to ensure genotyping call accuracy of the macro designed in the GeneMapperID v3.2 software (Applied Biosystems) and allele call. Using these collated data tables, CONVERT Ver.1.31

(Glaubitz, 2004) was used to calculate allele frequencies for each locus in each populations and to compile input files for use in other software tools.

2.4.1 Population statistics and neutrality testing

2.4.1.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg model describes a non-evolving population, in which allele frequencies do not change from generation to generation (Hardy, 1908). When allele frequencies do not change from generation to generation the population is considered to be in Hardy-Weinberg Equilibrium (HWE). Genetic drift, gene flow, mutation, non-random mating and natural selection are five factors which can influence evolution of a population or cause it to deviate from HWE.

In this study, each population was examined to establish whether the allelic variants were in HWE in ARLEQUIN version 3.0 (Excoffier et al., 2005). Analogous to HWE test, the exact test was further applied to data using in ARLEQUIN version 3.0 (Excoffier et al., 2005).

2.4.1.2 Neutrality testing

Tajima's D statistic (Tajima 1989a, 1989b, 1993) calculation was applied to each SNP locus using the ARLEQUIN version 3.0 software to determine which locus is under selection. Tajima's D is a statistical test that is able to distinguish between a DNA sequence evolving randomly ("neutrally") and one evolving under a non-random process, including directional selection or balancing selection, demographic expansion or contraction, genetic hitchhiking, or introgression. A randomly evolving

DNA sequence contains mutations with no effect on the fitness and survival of an organism. In other words the statistic tests the null hypothesis of mutation-drift equilibrium and constant population size.

2.4.1.3 Population diversity

Genetic diversity for each of the 18 populations in the study was calculated using allele frequencies for each locus analysed. To calculate the genetic diversity the standard diversity indices was selected in ARLEQUIN version 3.0 (Excoffier et al., 2005). The standard diversity index is proportional to the diversity within the population across the loci tested, i.e. the higher the index value (between 0 and 1), the greater the genetic diversity of that population.

2.4.2 Population affinity

To determine the genetic relationship of populations a pairwise comparison was conducted by calculating Fst distances in ARLEQUIN version 3.0 (Excoffier et al., 2005). The Fst distance matrix output of ARLEQUIN was use to view the spatial genetic relationship of populations through Principle Component Analysis (PCA) and cluster analysis in PAST v.1.54 (Hammer et al., 2001). Fst matrix was further used to construct population trees using the Neighbour Joining (NJ) method in PAST v.1.54 (Hammer et al., 2001).

To determine the distribution of variance among populations Analyses of Molecular Variance (AMOVA) was calculated in ARLEQUIN version 3.0 (Excoffier et al., 2005). The distribution of variance was determined in three levels:

- between individuals in a population

- between population within groups defined by geographic distribution and language
- between groups

2.4.3 Population structure

A widely used program called STRUCTURE is used to determine population structure using genotypic data from microsatellite DNA regions, restriction fragment length polymorphisms (RFLPs) or SNPs. The program assumes genotypes are in linkage equilibrium with one another in a population and that genotypes are in HWE proportion. In this study we used the program STRUCTURE Ver2.0 (Pritchard et al., 2000; Falush et al., 2003). The software uses allele frequencies in a Bayesian Clustering analysis to divide individuals into K number of clusters. The K is a prior set by the user, where STRUCTURE apportions genotypes of individuals into the K number of clusters with a certain probability. For example, if K=2 then STRUCTURE would divide the total genotypic variation for the entire sample then by statistical measure allocate individuals in the sample into either of the two clusters. For this analysis the admixed model was used along with default settings recommended for allelic data, those being for the correlated allelic frequency model. These settings assume that the populations share genetic traits where allelic frequencies are similar either due to shared ancestry or due to migration events. Since the number of clusters (K) is not known a number of runs were conducted from K=2 to K=10, to determine which K value apportions variation in populations with the greatest likelihood. A total of 10000 iterations with a burn-in phase of 1000 iterations were computed for each K cluster. Outputs of the inferred clustering was depicted in a bar chart output of the program, which was used to show the individual variation apportion for each individual for each K run. The program further outputs the average inferred clustering for each population. These population averages were

used as priors for the program Distruct Ver1.1 (Rosenberg, 2004), where the variation apportioned per population for each K run was depicted in a bar chart.

CHAPTER 3 RESULTS

This chapter describes the results that emerged from this study. The sections include the sample populations of the study and comparisons to other populations described in the comparative data collated (Appendix B). Study population labels used in this chapter correspond to the population codes that are listed Table 1. Firstly a brief discussion of the results of the techniques used to detect mutations is given. Secondly a description of the results obtained for mutations are provided. Thirdly trends within the study populations and published data are highlighted. Fourth the genetic affinities with respect to these markers are examined to describe the genetic histories of populations.

3.1. Mutation detection

The result of each run was analyzed using GeneMapperID v3.2 software (Applied Biosystems), where a macro was designed to score each SBE fragment automatically based on the fluorescent tag detected on the end of SBE fragments and the size of the SBE fragments. Figure 6 shows examples of electropherogram outputs by the GeneMapperID v3.2 software (Applied Biosystems). In Figure 6: sample (A) is heterozygous for CYP1A1*2A, CYP1A2*1C, CYP1A2*1F and GSTP1*Ile105Val; sample (B) is heterozygous for CYP1A1*2A, CYP1A1*2C and GSTP1*Ala114Val; sample (C) is heterozygous for CYP2A6*7 and CYP2D6*4; and sample (D) is heterozygous NAT2*14A. Populations in this study were all monozygotic for the wild-type allele for loci CYP2A6*8, CYP2D6*3A, CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I), therefore examples presented in Figure 6 have one peak each.

The Figure 6 also shows an additional marker than those mentioned above, the second peak from the left end of the electropherogram, which was designed to detect the NAT1*10 mutation (rs1057126). Unfortunately, following genotyping and checking of data it was discovered that the SBE primer designed for this marker was three nucleotide positions off in detecting the mutation. In reviewing reference sequences (extract April 2008) used to design the probe it was evident that the sequence used contained three bases less than that found in the most recent update (NCBI, ss279716933 last update 16 December 2010). The region of this SNP contains six AAT repeat sequences, whereas the reference sequence in 2008 contained only five. It is possible after the most recent realignment of the human DNA sequences submitted to NCBI database the reference sequence had changed to include the one AAT repeat, hence the error in the design of the SBE probe used. Due to this error all results for this marker were not included in the analysis for this study.

The large deletion mutations of the *GSTM1* and *GSTT1* genes, called GSTM1*0 and GSTT1*0 null mutations respectively, were genotyped using a multiplex reaction described earlier in section 2.3.2. The products of these reactions were resolved by electrophoresis on 2% Agarose gel. Images of these Agarose gels were captured and samples individually scored based on the bands visible from the gel images (Figure 7). The existence of a GSTM1 band or a GSTT1 band is an indication of the absence of GSTM1*0 and GSTT1*0, respectively. A control band from the β-Globin gene was included as a positive control in the event an individual contained deletions in both *GSTM1* and *GSTT1* genes, thereby distinguishing a true failed PCR from samples that contained both null mutations.

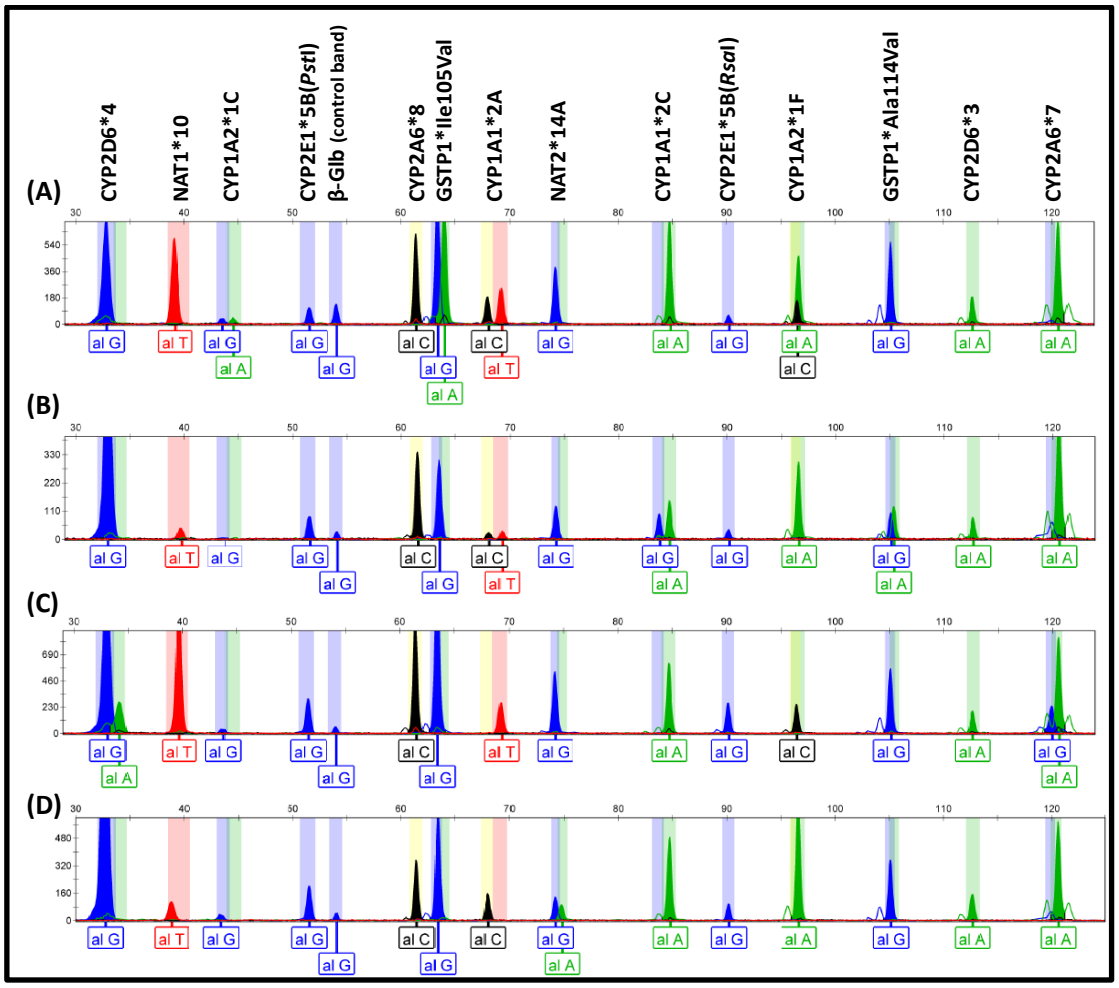

Figure 6 Examples of electropherogram results obtained during genotyping of SNP markers using SBE technique.

- SNP Marker names are labelled above the first electropherogram, where the peak position corresponds to the marker label above.
- Numbers on the X-axis of each electropherogram represent the fragment size as calculated using the internal size standard.
- Numbers on the Y-axis of each electropherogram represents peak heights
- Below each peak in the electropherograms is/are the allele/s calls generated by the genotyping software used and the macro developed for this purpose.
- The highlights in the background of peaks are the bin regions allocated for each marker in the macro, within which alleles could be scored. Colour highlights correspond to the nucleotide base call for the respective allele.
- Electropherogram (A) shows heterozygous loci CYP1A1*2A, CYP1A2*1C, CYP1A2*1F and GSTP1*Ile105Val; electropherogram (B) shows heterozygous loci CYP1A1*2A, CYP1A1*2C and GSTP1*Ala114Val; electropherogram (C) shows heterozygous loci CYP2A6*7 and CYP2D6*4; and electropherogram (D) heterozygous for NAT2*14A.

Due to the nature of the multiplex reaction it is not possible to distinguish between individuals who were heterogygous at these loci, therefore in the case of a band being present the sample was genotyped as not having the deletion mutation. Only where a band is absent can it be determined that an individual is homozygous for the respective null allele. An example of the results for this multiplex reaction can be found in Figure 7, where lane 1, 4, 5 and 7 are examples of individuals who were scored as having the wild-type genotype for both loci; lane 2 and 6 contains only the control band and thereby would have been scored as having both GSTM1*0 and GSTT1*0 genotypes; lane 8 does not have the GSTT1 band and thereby scored as having the GSTT1*0 genotype; lane 9 does not have the GSTM1 band and thereby scored as having the GSTM1*0 genotype; and lane 3 a 1 Kb Plus DNA Ladder (Invitrogen™) size standard.

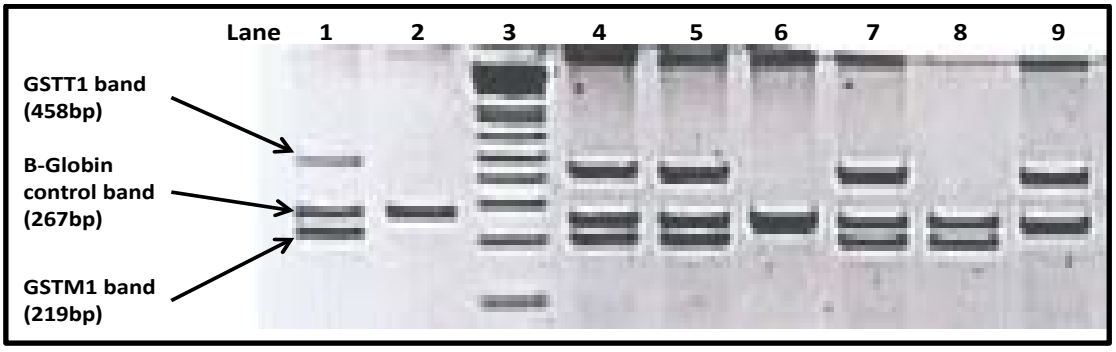

Figure 7 Gel picture showing results of GSTM1 and GSTT1 multiplex reaction

PCR products in the gel image above were resolved on a 2% Agarose gel
Represented in the figure above:

- Lanes 1, 4, 5 and 7 - are examples of individuals who were scored as having the wild-type genotype for both loci;
- Lanes 2 and 6 contains only the control band and thereby would have been scored as having both GSTM1*0 and GSTT1*0 genotypes;
- Lane 8 does not have the GSTT1 band and thereby scored as having the GSTT1*0 genotype;
- Lane 9 does not have the GSTM1 band and thereby scored as having the GSTM1*0 genotype;
- lane 3 a 1 Kb Plus DNA Ladder (Invitrogen™) size standard.

3.2. Phase I metabolizing genes

The following section describes the allele frequencies found for CYP1A1*2A and *2C; CYP1A2*1C and *1F; CYP2A6*7 and *8; CYP2D6*3A (2549delA) and CYP2D6*4(1846G>A); and CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I) polymorphisms. These polymorphisms are found in phase I metabolism genes described earlier. Data for these polymorphisms is presented in Table 5 and Figure 8 to Figure 17. Table 5 contains the frequencies of the mutant alleles identified in the study along with the total sample size tested. Comparative data for these polymorphisms can be found in Appendix B, Tables A to E. Current study and comparative data for these polymorphisms were collated in bar charts to illustrate the trends of these polymorphisms within Africa and globally, as shown in Figure 8 to Figure 17. Where more than one study reported findings for a mutation in that population the average frequency was used in the collation of bar charts.

3.2.1. CYP1A1*2A allele frequencies

In sub-Saharan Africa CYP1A1*2A is present in all populations, where Bantu-speakers display similar proportions of the allele varying from 26.6% in ZAM to 13.6% in UGA (Table 5 and Figure 8). The frequency is highest among the San-speakers (XUN, 39.6%; JU, 30.8%; KHWE, 28.6%) with lower frequencies in the Khoe-speakers (NAMA, 10.5%), a possible indication of the mutation's existence in Africa before human migration out of Africa. Pygmies were found to have similar high frequencies of the mutation as that found in Bantu-speakers (CAR_P, 25.9%). CYP1A1*2A allele is found in SAC at a frequency of 20.6%. Compared with parental populations, Khoe-San and SAW this appears to be intermediate in frequency. The SAI have the mutation present at 33%, much higher than that found in India. The

higher frequency in SAI is possibly due to founder events following the migration of Indians to South Africa.

The mutation tends to decrease further north outside sub-Saharan Africa towards the Middle East and Europe, becoming most prevalent in Asia especially the far-East (Figure 8 and comparative data in Appendix B, Table A). African-Americans, similar to their African relatives contain the mutation at a relatively high frequency (22.0%). In contrast, Caucasian Americans (9.0%) have similar low frequency for the mutation as their European relatives. Costa Ricans (31.4%) and Chileans (37.0%) too contain large frequencies for CYP1A1*2A allele. Overall, the data available tends to indicate that the mutation can be found in all populations globally at varying frequencies.

3.2.2. CYP1A1*2C allele frequencies

CYP1A1*2C is near absent in sub-Saharan Africa (Table 5 and Figure 9). The mutation is found in Ivory Coast (38.3%), Ghana (8.0%), HIM (1.3%), and MOZ (0.9%), where it is likely that individuals in these populations acquired the mutation from non-African descendants in the recent past (Figure 9 and comparative data in Appendix B, Table A). A likely corroboration of this aspect can be found in the SAC where the mutation is found at 5.3%. Given the ancestral history of the SAC, the near absence of the allele in neighbouring indigenous populations, the CYP1A1*2C would more likely have been introduced by parental populations SAI (13.8%) and SAW (9.1%).

CYP1A1*2C increases in frequency from North Africa (Egypt, 27%), towards the Middle East, to being most frequent in Asia. The mutation is found at appreciably lower frequencies in Europe (ranging between 2% and 5%). Consistent with their

heritage Caucasoid populations of North America contain low frequencies of the mutation. African-Americans have the mutation as well, possibly due to the introduction from their European descendants similar to that observed for the SAC. In South America, the mutations is present at high frequencies in Mexico (50%) and Chile (32%), to lower a frequency in Brazil (16%), possibly due to the sample in Mexico and Chile mostly consisting of Amerindian populations (Figure 9 and comparative data in Appendix B, Table A).

3.2.3. CYP1A2*1C allele frequencies

During raw data checks of SBE runs for this marker it was apparent that the reaction did not work equally well for some populations where the missing data ranged from 8.1% (DAMA) to 96.3% (JU), where average missing data was 49%. Due to low peak heights of this marker not all typings could be considered as true allele scoring for this locus and therefore were not included in the final dataset. Figure 6 (sample B) shows an example of such a sample where the amplification of this locus was very low and the result could not be considered as true. Due to the amplification success in some populations and not others it is likely that the lack of success in typing this locus was sample dependent. For this reason this study cannot confirm the true frequency profiles of this locus in the study populations. Evident though, the mutation is present in most in sub-Saharan African populations (Table 5 and Figure 10). From comparative data available the allele tends to be most frequent in far-East Asia (Japan, 26%), to 7% in India and Egypt (Figure 10 and comparative data in Appendix B, Table B). The mutation is found at appreciably lower frequencies in Europeans ranging from 1% to 4%, and it is therefore likely that SAW could have this mutation.

Table 5 SNP loci mutant allele frequencies only – current study population

African region	Southern Africa											
Country of sample	South Africa		South Africa & Nambia		South Africa & Nambia		South Africa		South Africa		South Africa	
Population Code ¥	SEB		XUN		KHWE		SAC		SAI		SAW	
Locus	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)
CYP1A1*2A rs4646903T>C	70	22.1	67	39.6	49	28.6	85	20.6	103	33.0	42	14.3
CYP1A1*2C rs1048943A>G	70	0.0	69	0.7	49	0.0	85	5.3	105	13.8	44	9.1
CYP1A2*1C rs2069514G>A	33	27.3	29	31.0	30	31.7	42	3.6	58	3.5	10	0.0
CYP1A2*1F rs762551C>A	70	59.3	69	66.7	49	61.2	85	68.8	105	62.9	44	71.6
CYP2A6*7 rs5031016A>G	70	0.7	69	8.0	49	3.1	85	5.9	105	5.7	44	3.4
CYP2A6*8 rs28399468C>A	70	0.0	69	0.0	49	0.0	85	0.0	105	0.0	44	0.0
CYP2D6*3 rs35742686delA	67	0.0	61	0.0	47	0.0	78	0.0	103	0.0	34	0.0
CYP2D6*4 rs3892097G>A	70	1.4	69	0.0	49	3.1	85	5.3	104	5.3	43	10.5
CYP2E1*5B (PstI) rs3813867G>C	69	0.0	66	0.0	49	0.0	83	0.0	103	0.0	37	0.0
CYP2E1*5B (RsaI) rs2031920G>A	68	0.0	61	0.0	44	0.0	75	0.0	75	0.0	34	0.0
GSTP1*Ile105Val rs1695A>G	70	55.0	69	72.5	49	57.1	85	48.2	105	27.6	44	38.6
GSTP1*Ala114Val rs1138272G>A	70	1.4	69	0.0	49	0.0	85	4.7	105	6.7	44	14.8
NAT2*14A rs1801279G>A	70	8.6	68	4.4	49	5.1	85	5.3	105	0.0	44	0.0

¥ For population name and ethnicity refer to table 1
N represents the total sample size

Table 5 continued… SNP loci mutant allele frequencies only – current study population

African region	Southern Africa											
Country of sample	Botswana		Namibia		Namibia		Namibia		Namibia		Namibia	
Population Code ¥	JU		NAMA		DAMA		HER		HIM		OVA	
Locus	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)
CYP1A1*2A rs4646903T>C	26	30.8	19	10.5	37	23.0	78	25.6	40	22.5	60	21.7
CYP1A1*2C rs1048943A>G	27	0.0	19	0.0	37	0.0	78	0.0	40	1.3	60	0.0
CYP1A2*1C rs2069514G>A	1	100.0	9	5.6	34	17.7	39	37.2	36	33.3	23	17.4
CYP1A2*1F rs762551C>A	27	70.4	19	71.1	37	39.2	78	61.5	40	50.0	60	52.5
CYP2A6*7 rs5031016A>G	27	11.1	19	0.0	37	0.0	78	0.6	40	1.3	60	0.0
CYP2A6*8 rs28399468C>A	27	0.0	19	0.0	37	0.0	78	0.0	40	0.0	60	0.0
CYP2D6*3 rs35742686delA	22	0.0	18	0.0	37	0.0	77	0.0	39	0.0	60	0.0
CYP2D6*4 rs3892097G>A	27	0.0	19	0.0	37	2.7	78	3.2	40	0.0	60	2.5
CYP2E1*5B (PstI) rs3813867G>C	26	0.0	19	0.0	37	0.0	76	0.0	40	0.0	60	0.0
CYP2E1*5B (RsaI) rs2031920G>A	15	0.0	17	0.0	34	0.0	77	0.0	39	0.0	60	0.0
GSTP1*Ile105Val rs1695A>G	27	74.1	19	65.8	37	56.8	78	55.1	40	46.3	60	45.8
GSTP1*Ala114Val rs1138272G>A	27	0.0	19	0.0	37	0.0	78	1.3	40	0.0	60	0.0
NAT2*14A rs1801279G>A	27	0.0	19	5.3	37	1.4	78	10.3	40	13.8	60	10.0

¥ For population name and ethnicity refer to table 1
N represents the total sample size

Table 5 continued… SNP loci mutant allele frequencies only – current study population

African region	Eastern Africa						Middle Africa					
Country of sample	Mozambique		Zambia		Uganda		Central African Republic		Central African Republic		Democratic Republic of Congo	
Population Code ¥	MOZ		ZAM		UGA		CAR_U		CAR_P		DRC	
Locus	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)	N	Mutant allele (%)
CYP1A1*2A rs4646903T>C	52	20.2	175	26.6	224	13.6	177	22.3	81	25.9	160	21.9
CYP1A1*2C rs1048943A>G	54	0.9	175	0.0	227	0.0	179	0.0	91	0.0	171	0.0
CYP1A2*1C rs2069514G>A	40	10.0	137	21.2	160	18.1	87	16.7	25	34.0	33	27.3
CYP1A2*1F rs762551C>A	54	50.0	175	54.6	227	46.9	179	56.4	91	76.4	170	53.8
CYP2A6*7 rs5031016A>G	54	2.8	175	1.7	227	3.3	179	11.7	91	3.9	171	23.7
CYP2A6*8 rs28399468C>A	54	0.0	175	0.0	227	0.0	179	0.0	91	0.0	171	0.0
CYP2D6*3 rs35742686delA	53	0.0	172	0.0	226	0.0	164	0.0	90	0.0	129	0.0
CYP2D6*4 rs3892097G>A	54	2.8	175	3.7	227	3.1	179	3.1	91	3.3	171	7.0
CYP2E1*5B (PstI) rs3813867G>C	52	0.0	175	0.0	225	0.0	171	0.0	82	0.0	121	0.0
CYP2E1*5B (RsaI) rs2031920G>A	54	0.0	170	0.0	223	0.0	165	0.0	72	0.0	123	0.0
GSTP1*Ile105Val rs1695A>G	54	45.4	175	42.6	227	44.9	179	43.9	91	36.8	169	43.8
GSTP1*Ala114Val rs1138272G>A	54	0.9	175	0.0	227	0.7	179	0.0	91	0.0	171	0.3
NAT2*14A rs1801279G>A	54	9.3	175	11.1	227	7.3	179	6.2	91	4.4	171	9.1

¥ For population name and ethnicity refer to table 1
N represents the total sample size

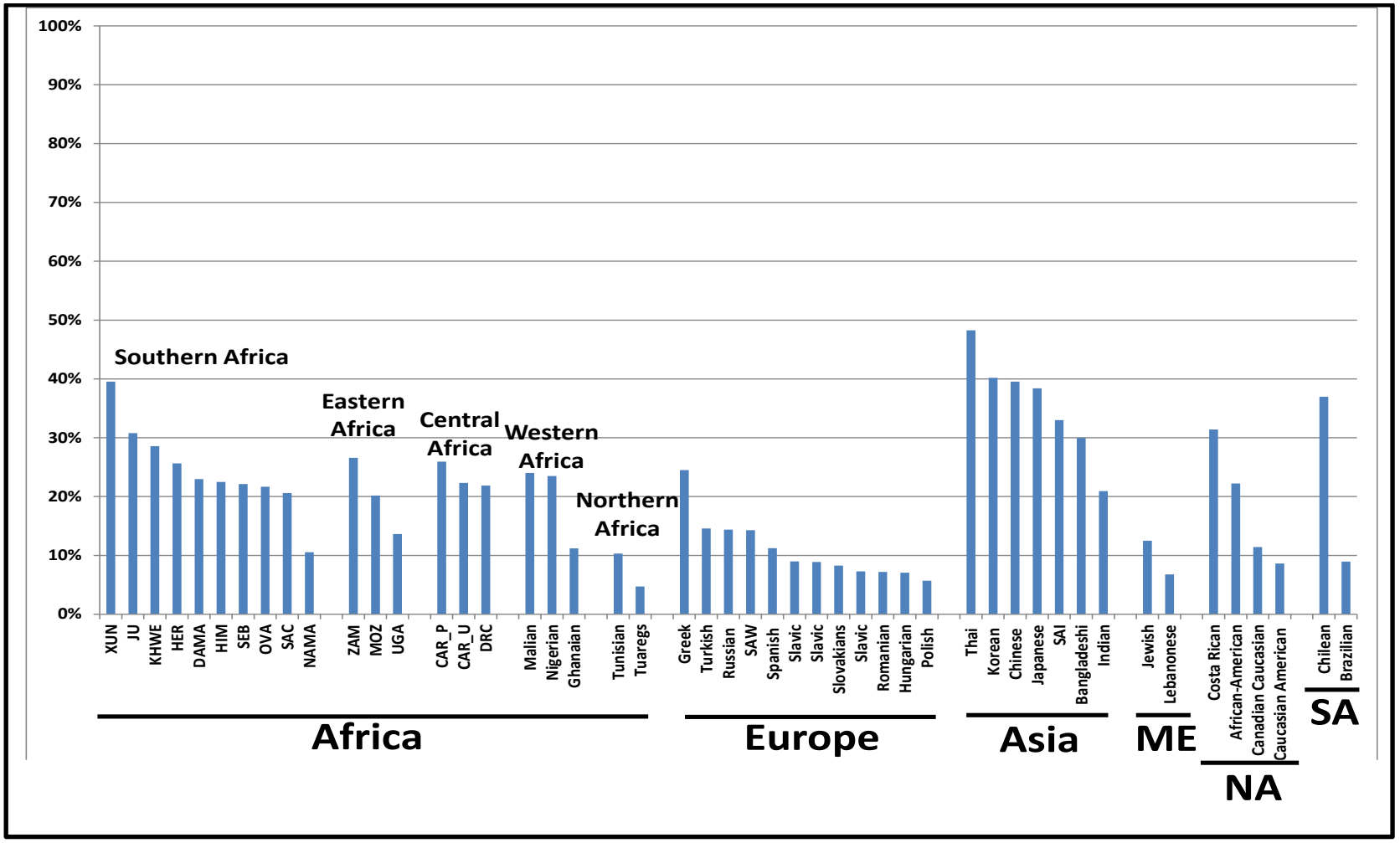

Figure 8 Bar chart showing CYP1A1*2A allele proportions worldwide

NB: ME = Middle East; NA = North America; SA = South America

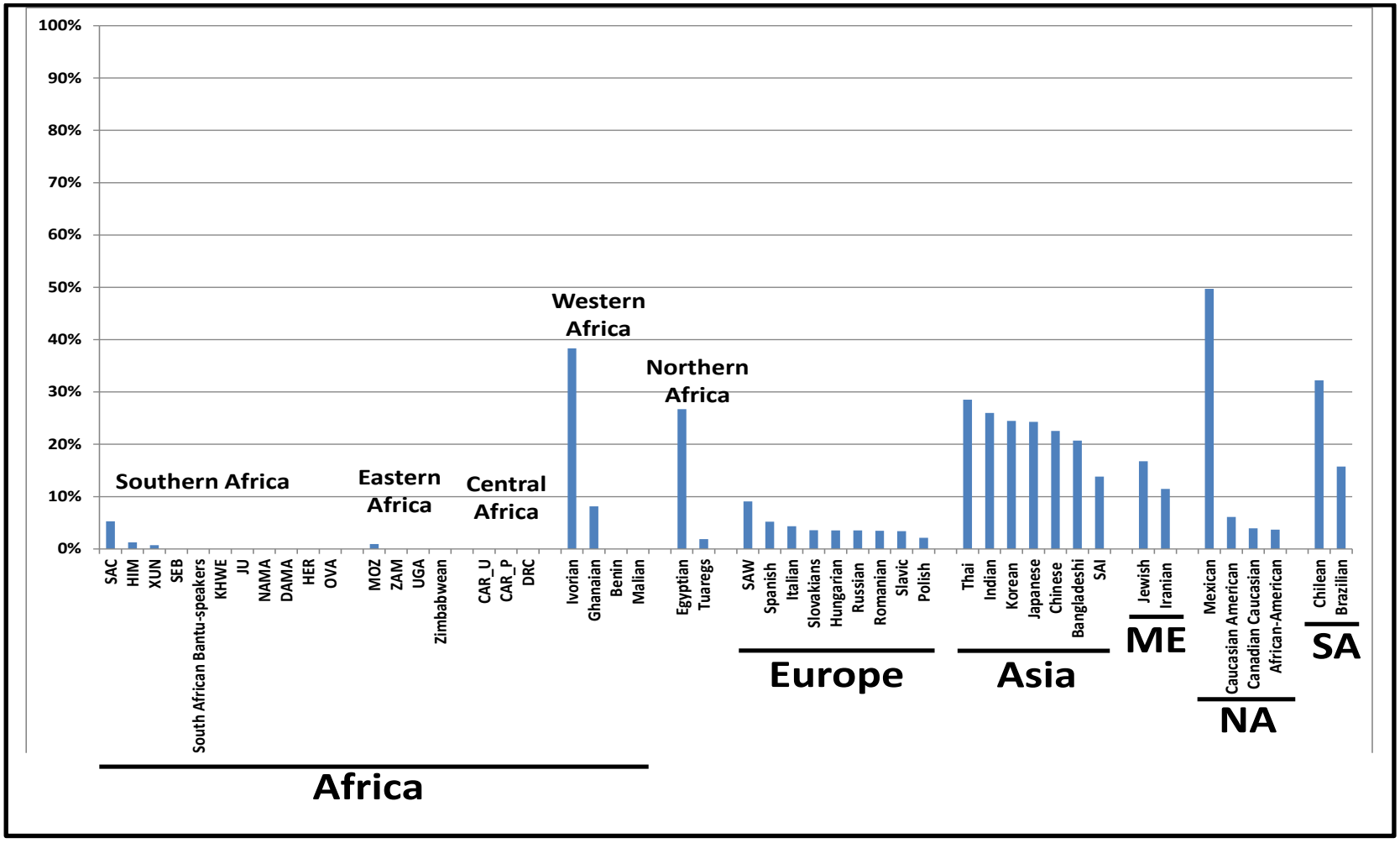

Figure 9 Bar chart showing CYP1A1*2C allele proportions worldwide

NB: ME = Middle East; NA = North America; SA = South America

3.2.4. CYP1A2*1F allele frequencies

The CYP1A2*1F allele is found throughout the globe (Table 5 and Figure 11 with comparative data in Appendix B, Table B). In sub-Saharan Africans the mutation is present at frequencies greater than 40% (Table 5 and Figure 11). There is no apparent trend in the distribution of this mutation throughout Africa. Its high prevalence in Khoe-San and CAR_P populations indicates that the allele had evolved in Africa before humans migrated out of Africa. Similar as in Africa, the mutation is present at high frequencies in Europe, Asia, North America and South America to varying degrees. The CYP1A2*1F allele in the SAI (62.9%) is greater than that found in India (44%), where the difference can be attributed to the Indian sample mainly consisting of North Indians and the SAI being migrants from North, South and eastern India.

3.2.5. CYP2A6*7 allele frequencies

CYP2A6*7 allele is present among all sub-Saharan African populations investigated here, with the exception of the NAMA, DAMA and OVA (Table 5 and Figure 12).. The allele is most frequent in Central Africa to being low in southern Africans. The allele is present in San populations (JU, 11.1%; XUN, 8%; and KHWE, 3.1%) and absent in Khoe (NAMA). Data from literature is scarce where no data was found for North Africa, Middle Eastern and South American populations. From the data available for this locus, the mutation is present in far-East Asians and absent in Caucasians and African Americans of Canada (Figure 12 and comparative data in Appendix B, Table C). While Indian and Swedish populations do not have the mutation, it is present in the SAI and SAW respectively. This indicates that the allele could be present in Indians and Europeans, likely at low frequencies.

3.2.6. CYP2A6*8 allele frequencies

The CYP2A6*8 allele is absent in sub-Saharan Africans (Table 5 and Figure 13). From current comparative data the mutation is present in far-East populations but absent elsewhere (Figure 13 and comparative data in Appendix B, Table C). The marker's exclusivity in Asia cannot be confirmed though as the data is underrepresented for several global populations. Further investigation is required within Africa and globally to understand the allele's distribution.

3.2.7. CYP2D6*3A (2549delA) allele frequencies

Data for this allele is shows that a single allele is present only in Tanzanians (0.3%) and absent in all other investigated African populations (Table 5 and Figure 14). Outisde Africa, CYP2D6*3A is absent in Asians; while present in Europeans, Iranians, and Caucasoid populations of North America (Figure 14 and comparative data in Appendix B, Table D). The allele is found in African-Americas (0.6%). Given the history of this population and the absence of this marker in Africans it is likely that the presence of this allele was contributed by European descendants. The mutation is found in North and South Americans, albeit at low frequency in Mexican (0.4%), American Indians (0.3%), Argentinians (2.5%), and Chileans (1.0%). Its absence in other American populations (Panamanian and Columbian) and the low prevalence in those listed above seems to indicate that the mutation is not indigenous to Native Americans, and rather was introduced by European settlers. The high proportion in Brazilians (13.8%) can be explained by the distribution of the sample where the authors reported the sample comprising mostly (86.4%) of those who phenotypically resembled Europeans (Silviera et al., 2010).

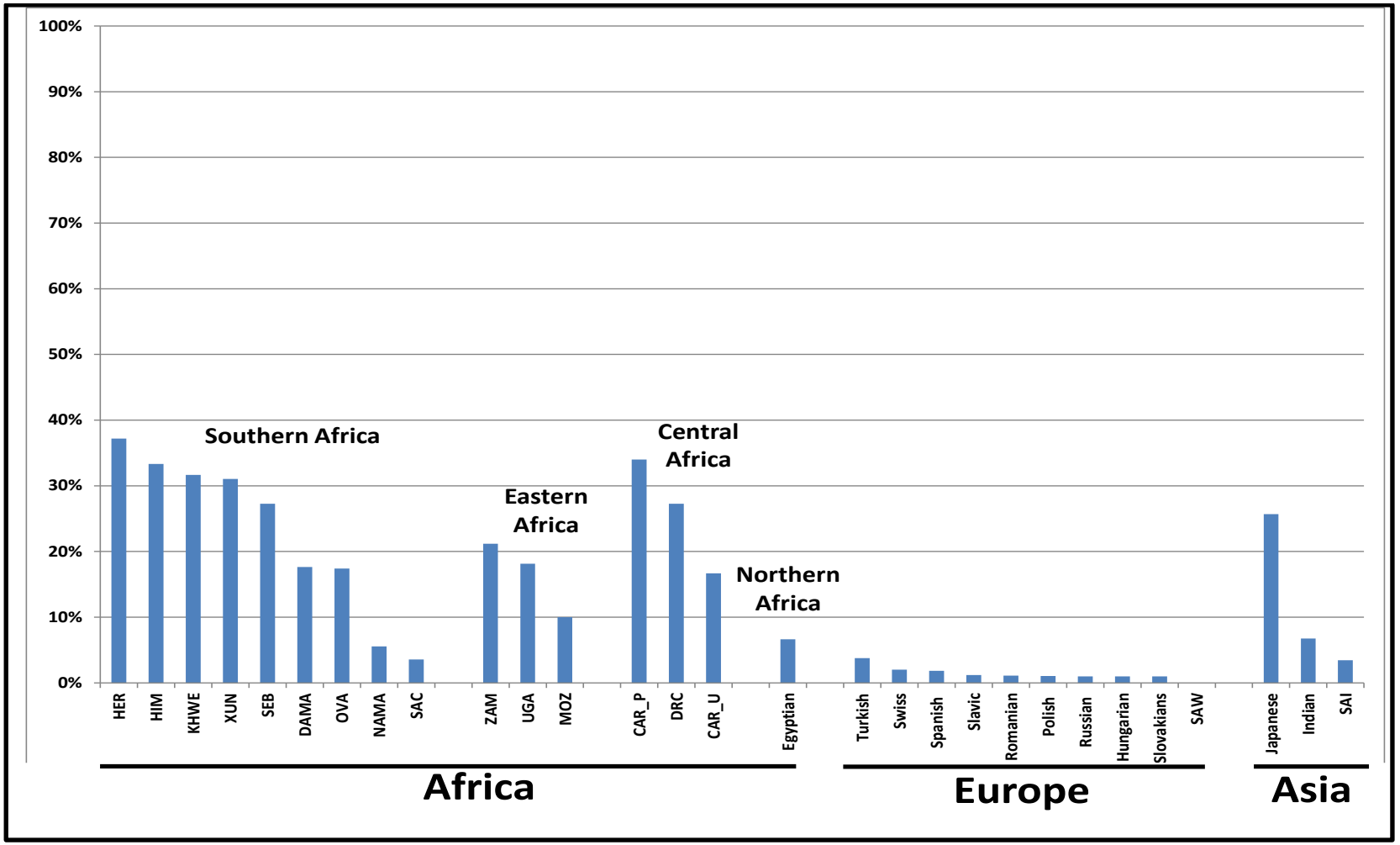

Figure 10 Bar chart showing CYP1A2*1C allele proportions worldwide

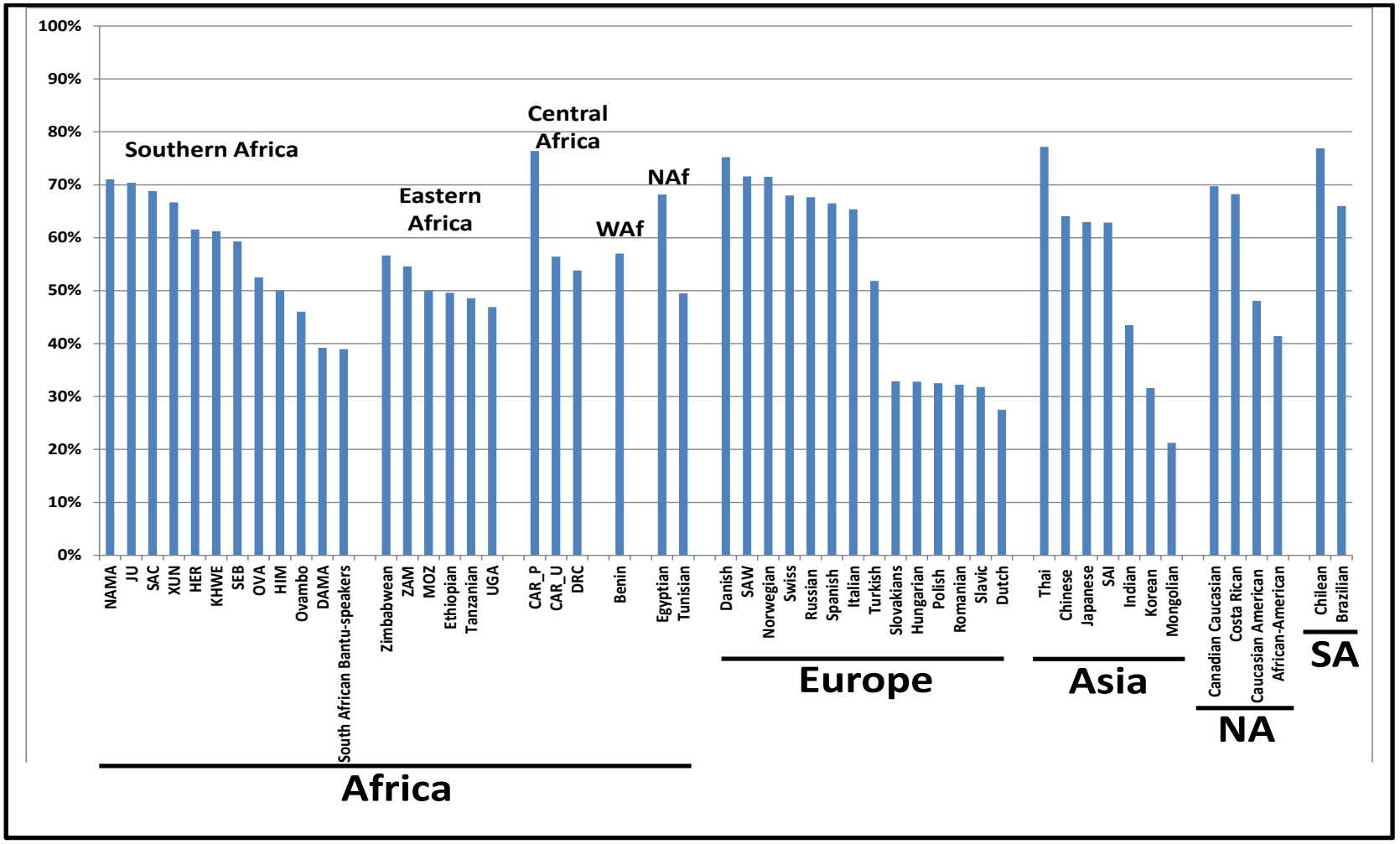

Figure 11 Bar chart showing CYP1A2*1F allele proportions worldwide
NB: WAf = Western Africa; NAf = Northern Africa; NA = North America; SA = South America

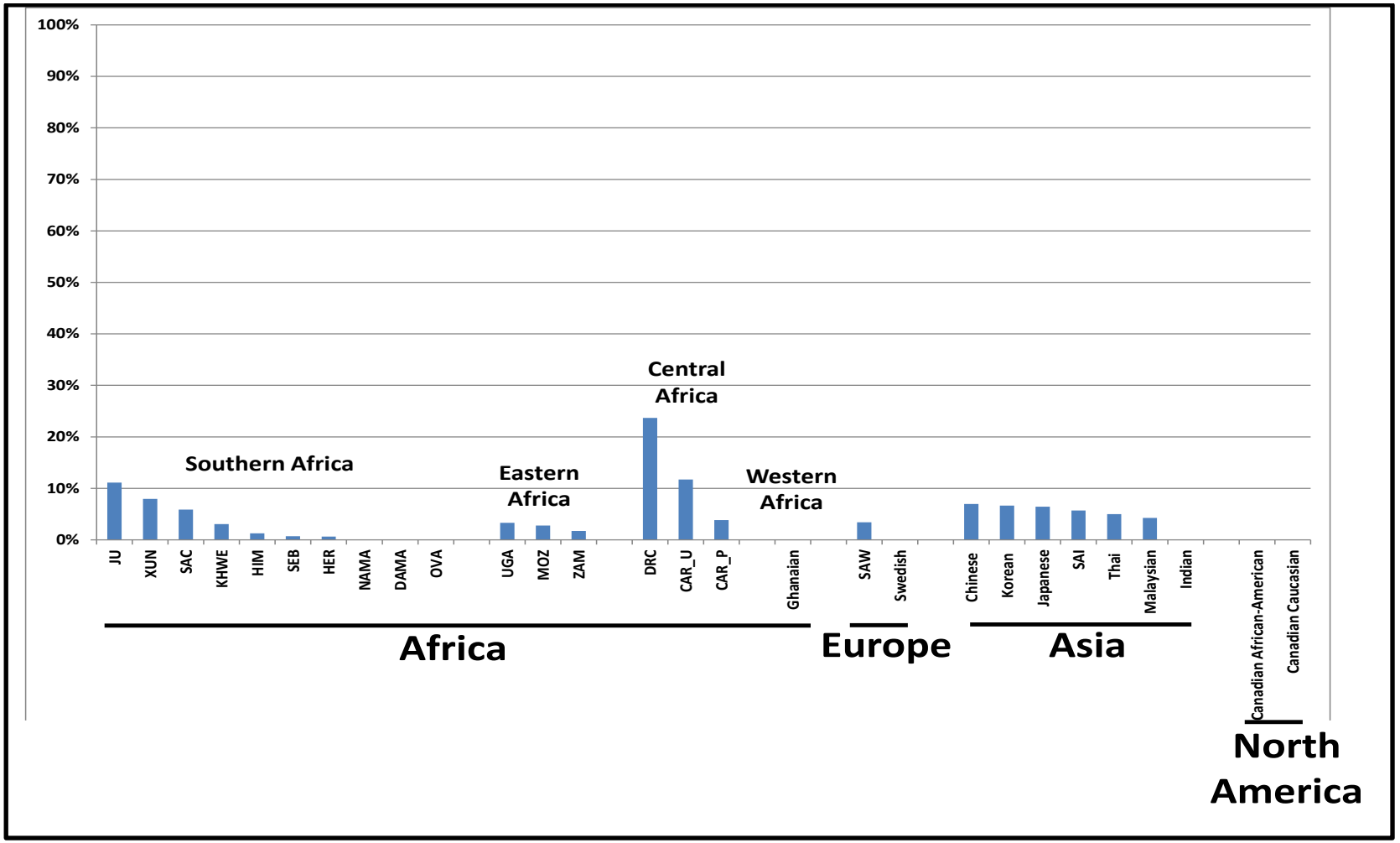

Figure 12 Bar chart showing CYP2A6*7 allele proportions worldwide

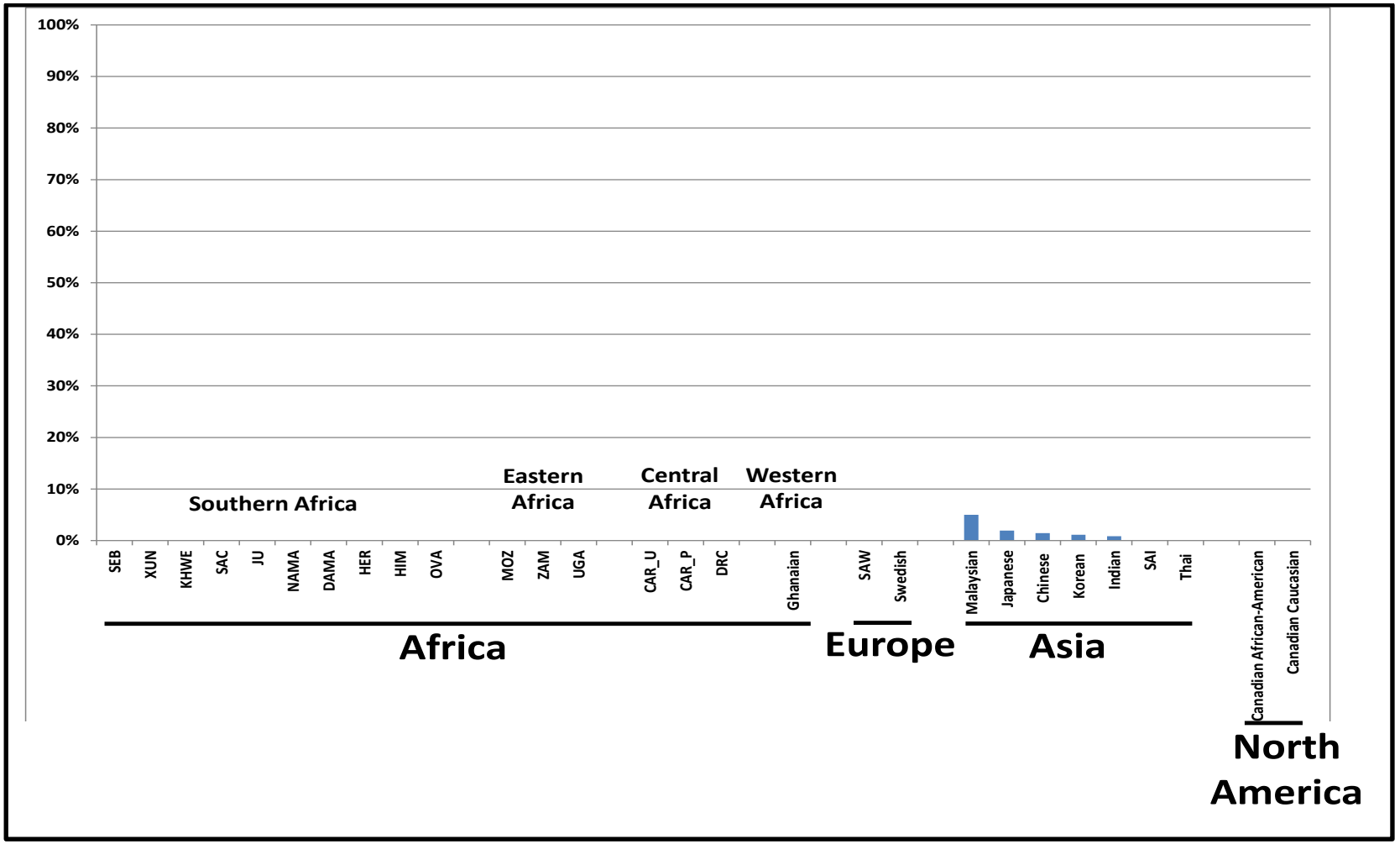

Figure 13 Bar chart showing CYP2A6*8 allele proportions worldwide

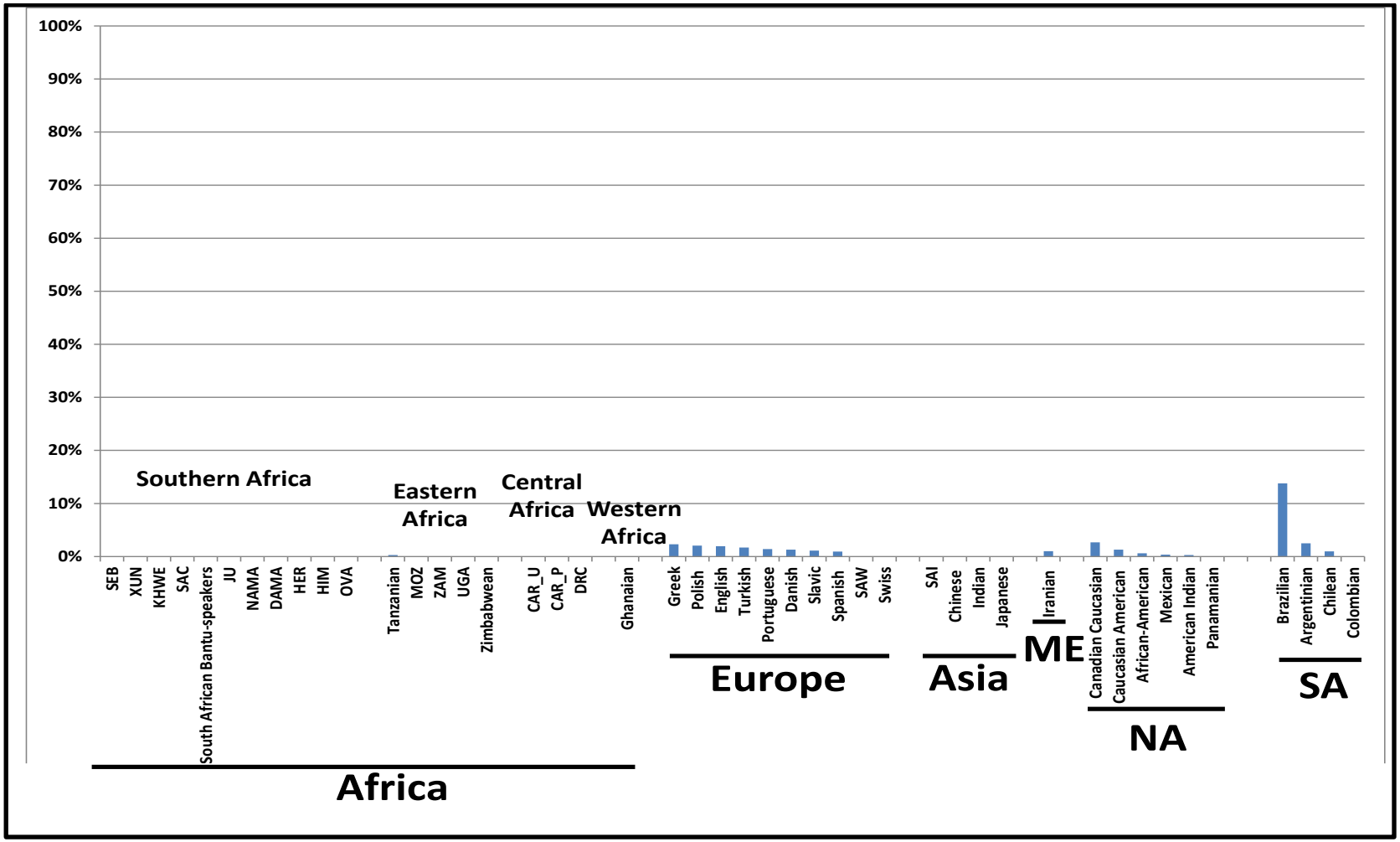

Figure 14 Bar chart showing CYP2D6*3A allele proportions worldwide
NB: ME = Middle East; NA = North America; SA = South America

3.2.8. CYP2D6*4 (1846G>A) allele frequencies

CYP2D6*4 is absent in all Khoe-San populations, except the KHWE (3.1%) (Table 5 and Figure 15). CYP2D6*4 is present in all Bantu-speakers including the CAR_P (ranging from 1% to 7%), where the frequency increases in the west (Ghana, 13%) and north (Tunisia, 9%) Africa. DAMA exhibit a similar proportion of the allele (2.7%) as in Bantu-speakers. This corroborates other findings that DAMA are biologically more closely related to Bantu-speakers, even though they speak Nama (Nurse et al., 1985; Soodyall and Jenkins, 1993; Soodyall et al., 2008; Schlebusch, 2010a). South African Bantu-speakers and South African Coloured populations of the current study and in literature all show that the allele is present at similar proportions as that found in other southern Bantu-speakers.

The allele is most frequent in Europeans and other European derived populations (Figure 15 and comparative data in Appendix B, Table D). CYP2D6*4 is found in Middle East, North American and South American populations. Given the high prevalence in Native Americans it is possible that the mutation was present in these populations prior to European settlement, when Europeans were introduced or a combination of both these reasons. The mutation is present in Indians (14%) and SAI (5.3%). In contrast to the trend seen in Asia for other locus investigated here, the mutation is near absent in far-East Asians. SAW and SAI contain the mutant allele at lower frequencies compared to parental populations, likely due to the founder events following their migration to South Africa.

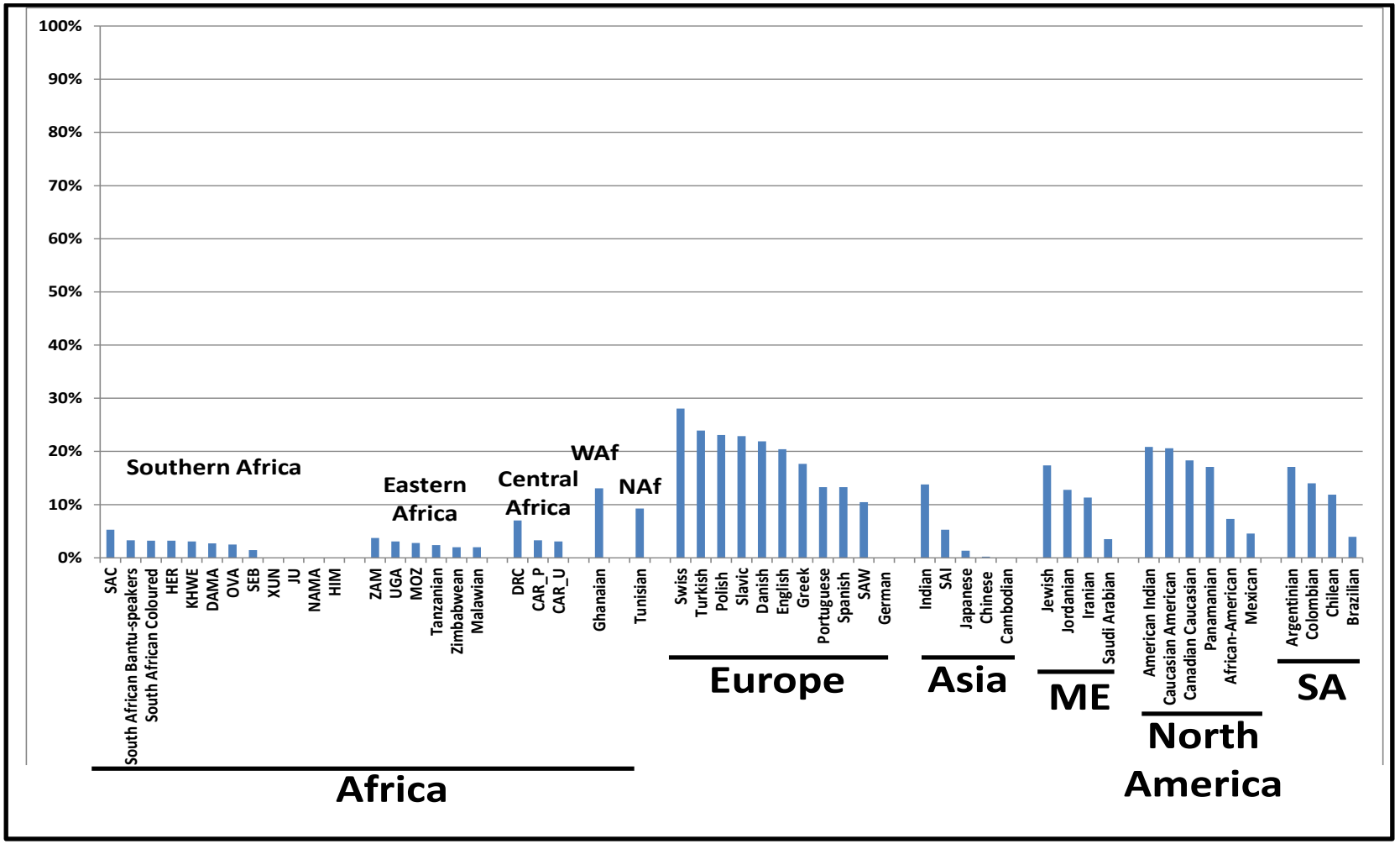

Figure 15 Bar chart showing CYP2D6*4 allele proportions worldwide
NB: WAf = Western Africa; NAf = Northern Africa; ME = Middle East; SA = South America

3.2.9. CYP2E1*5B(*Pst*I) allele frequencies

The CYP2E1*5B(*Pst*I) allele is absent in sub-Saharan Africans, with highest frequencies observed in far-East Asians, Chileans and Caucasoid populations of North America (Table 5 and Figure 16 with comparative data in Appendix B, Table E). The allele is present in Europe and India at low frequencies, and absent in Lebanon. Comparative data for the mutation was not available for most western Europeans. Given the migrant history of Europeans to North America, its presence in Caucasians from the United States of America and Canada indicates that it possibly exists in Western Europe. The allele's absence in SAW is surprising and is possibly due to the founder events of these populations.

3.2.10. CYP2E1*5B(*Rsa*I) allele frequencies

CYP2E1*5B(*Rsa*I) is absent in all sub-Saharan African populations (Table 5 and Figure 17 and comparative data in Appendix B, Table E), with the exception of a minute proportion found in South African Bantu-speakers from a literature source (2.7%, Chelule et al., 2006). The presence of nine CYP2E1*5B(*Rsa*I) alleles in this study is interesting as the mutation is absent in other southern African Bantu-speakers. This study's sample was collected in Kwa-Zulu Natal, different to the SEB sample of the current study. This difference in samples could indicate the introgression of the allele from European and/or Indian descendants. The mutation is further found in North Africa (Egypt, 1.7%).

Although there is limited comparative data for this allele compared to CYP2E1*5B(*Pst*I), CYP2E1*5B(*Rsa*I) appears to be similarly distributed globally as that observed for the CYP2E1*5B(*Pst*I) allele. This is not surprising as many studies have reported that the two loci are highly linked to one another (Hildesheim et al.,

1995; Gonzalez et al., 1998; Rodrigo et al., 1999; Kim et al., 2000; Sarmanova et al., 2001; Kim et al., 2004; Landi et al., 2005; Chen et al., 2007; Gemignani et al., 2007; Hsieh et al., 2007; Nishino et al., 2008; Wickliffe et al., 2011; Anderson et al., 2012; Huo et al., 2012). CYP2E1*5B(*Rsa*I) is found at low frequencies in Europeans and South Americans. In contrast to the trend seen for CYP2E1*5B(*Pst*I), CYP2E1*5B(*Rsa*I) is present in Lebanon at low frequencies (1.4%).

3.3. Phase II metabolizing genes

The following section describes the genotype frequencies of the GSTM1*0 and GSTT1*0 null mutations, and allele frequencies for GSTP1*Ile105Val and *Ala114Val; and NAT2*14A polymorphisms. These polymorphisms are found in phase II metabolism genes described earlier. Data for these polymorphisms are presented in Table 5 and Table 6, and in Figure 18 to Figure 23. Table 6 contains genotype frequencies obtained for the GSTM1*0 and GSTT1*0 null mutations for populations investigated in this study. Comparative data for GSTM1*0 and GSTT1*0 null mutations can be found in Appendix B, Table F. Table 5 contains the frequencies of both wild-type and mutant alleles of GSTP1*Ile105Val, GSTP1*Ala114Val and NAT2*14A polymorphisms. Comparative data for GSTP1*Ile105Val, GSTP1*Ala114Val and NAT2*14A polymorphisms can be found in Appendix B, Tables G and H. Current study and comparative data for these polymorphisms were collated in bar charts to illustrate the trends of these polymorphisms within Africa and globally, as shown in Figure 18 to Figure 23. Where more than one study reported findings for a mutation in that population the average frequency was used in the collation of bar charts.

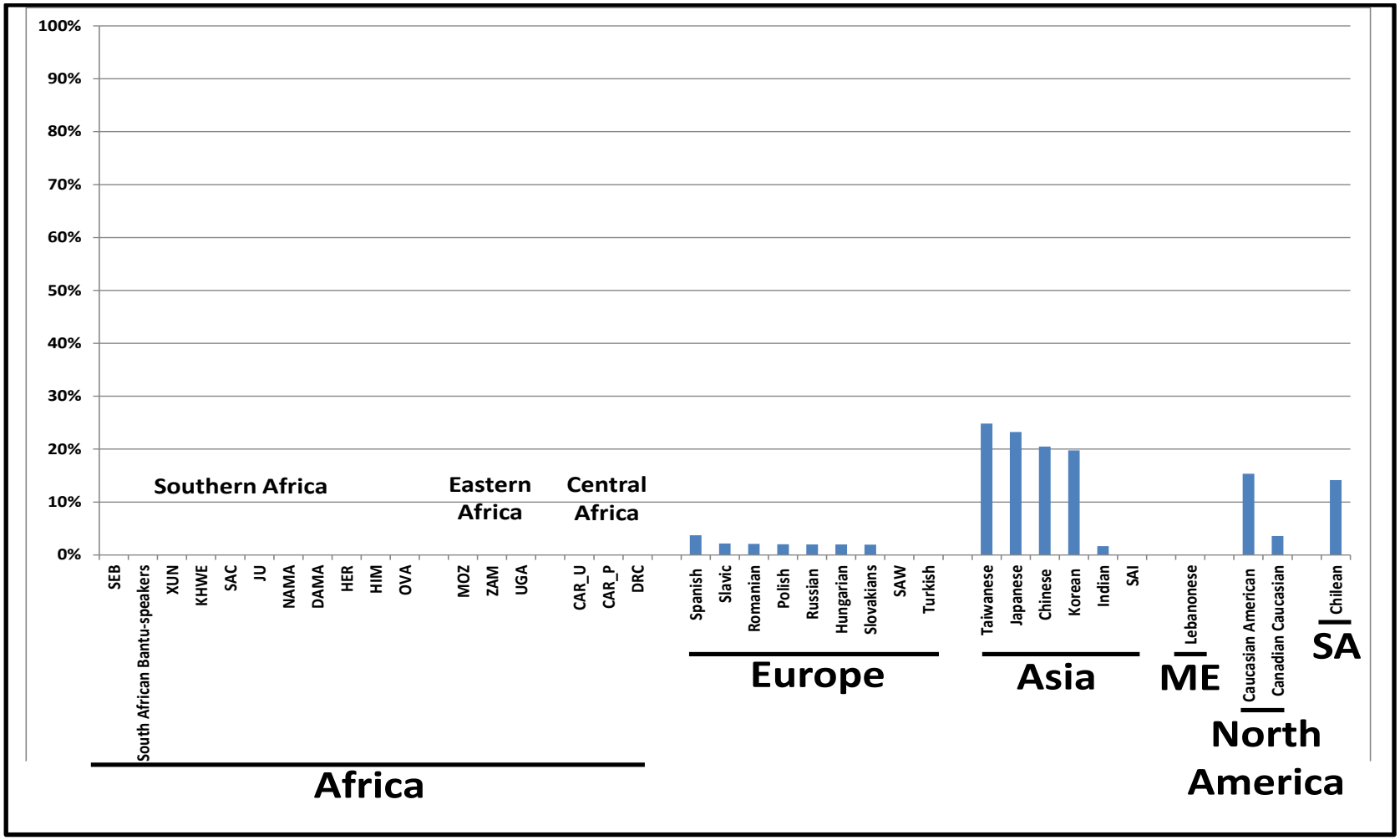

Figure 16 Bar chart showing CYP2E1*5B(*Pst*I) allele proportions worldwide
NB: ME = Middle East; SA = South America

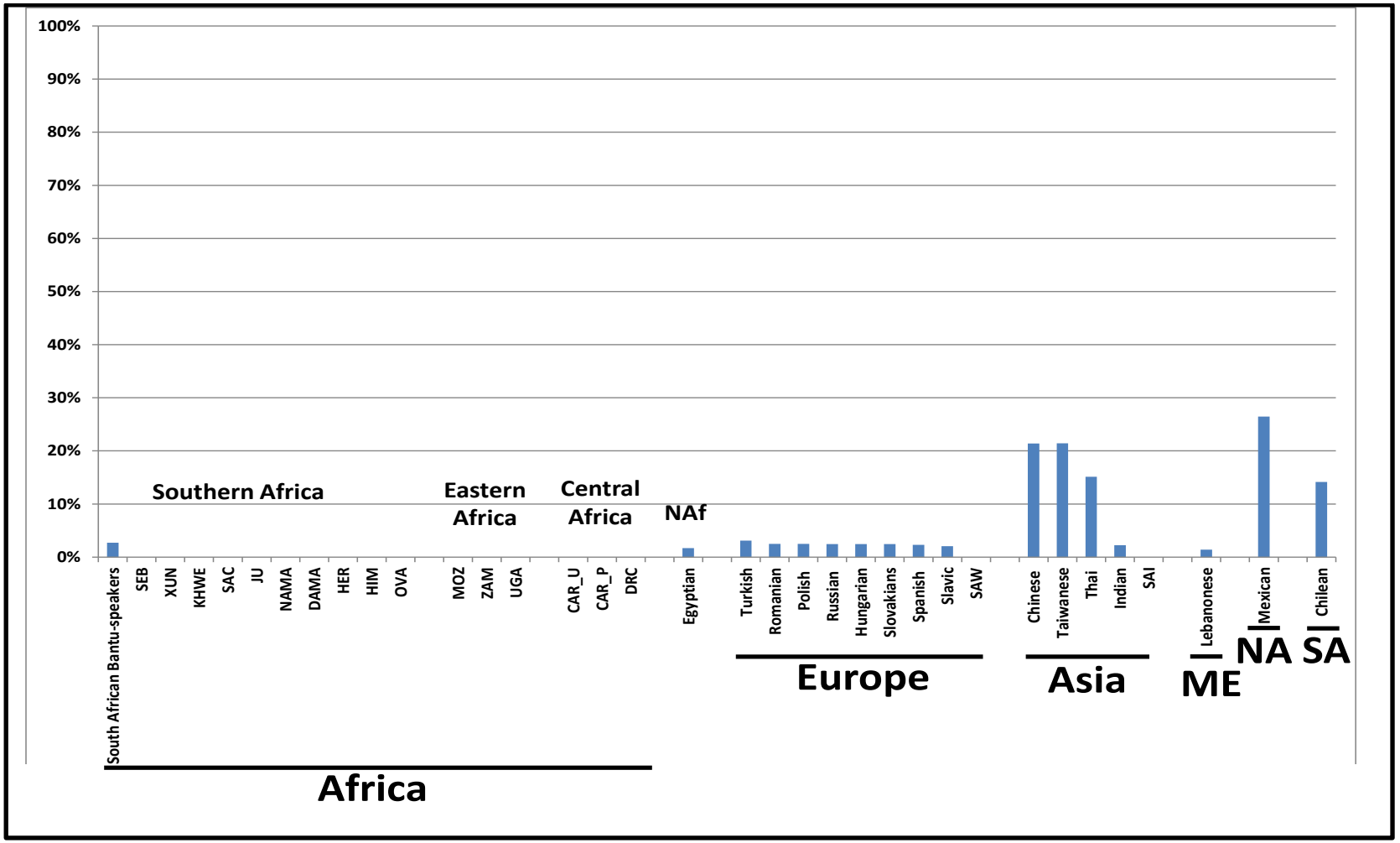

Figure 17 Bar chart showing CYP2E1*5B(*Rsa*I) allele proportions worldwide
NB: NAf = Northern Africa; ME = Middle East; NA = North America; SA = South America

3.3.1. GSTM1*0 genotype frequencies

In sub-Saharan Africa, the GSTM1*0 mutation is present in almost all populations at varying frequencies (Table 6 and Figure 18). Khoe-San populations, with the exception of KHWE (13%), contain low frequencies of the GSTM1*0 genotype, to being absent in JU. Southern Bantu-speakers contain GSTM1*0 at frequencies ranging from 22.6% (OVA) to 5.4% (DAMA). Its distribution in Bantu-speakers appears to be similar in eastern, western and Central Africa, ranging from 36.1% (Ivory Coast) to 2.8% (Cameroon). In South African Coloured group, GSTM1*0 was found at a frequency of 22.2%, somewhat intermediate between Africans and Europeans.

The mutation's frequency increases from the south to North Africa and continues out of Africa being most frequent in Europe, Asia and the Middle East (Figure 18 with comparative data in Appendix B, Table F). In the Middle East the GSTM1*0 genotype ranges between 37.4% (Iran) to 53.4% (Saudi Arabia). Middle East's high prevalence is not surprising considering that this region borders Europe where the frequency ranges from 66.1% (France) to 49.1% (Italy). In Asia, GSTM1*0 is low in Indians (30%), becoming more frequent towards the far-East. The mutation is further present in North and South Americans, where the frequency is largest in Caucasoid populations of North America (Canadian Caucasians, 56.5%; Caucasian Americans, 53.5%).

3.3.2. GSTT1*0 genotype frequencies

The mutation is found in all Khoe-San populations investigated at varying frequencies from 41.3% (KHWE) to 33.8% (XUN) (Table 6 and Figure 19). Among southern Bantu-speakers the GSTT1*0 genotype ranged between 41.0% (HIM) to 16.4% (ZAM). Further north there appears to be no distinct trend in the distribution of GSTT1*0 among Central, eastern, western and Northern Africans (Figure 19, with comparative data in Appendix B, Table F). In contrast to most

Bantu-speaking populations investigated in the study, the GSTT1*0 genotype is present at a very low frequency in the CAR_P (5.9%). GSTT1*0 was found at a frequency of 21.7% in SAC, this is somewhat intermediate to that found in parental populations. The frequency in SAC is much lower than the South African Coloured group (SAC[1], 57%) described in Adams et al. (2003). This Coloured group was sampled in the Western Cape of South Africa, different to those in this study who were sampled primarily in Johannesburg. These two groups do have different histories where it is known that the Coloured groups of Western Cape predominantly have been born from the Khoe-San natives and European migrants (Schlebusch, 2010a).

Globally, the GSTT1*0 genotype is found in Europe, Middle East, Asia and the Americas. In Europe, with the exception of Greece (43.3%), the mutation is found at lower frequencies than that found in Africa (Figure 19, with comparative data in Appendix B, Table F). The GSTT1*0 genotype is found in far-East Asians at similar high frequencies as that in Africans, ranging from 50.9% (Koreans) to 25.6% (Mongolians). The mutation is further found in the Middle East ranging between 39% (Saudi Arabia) and 19.2% (Iran). With the exception of Greenland (46%), GSTT1*0 is found at relatively lower frequencies. This is possibly due to the population being primarily small and isolated as compared to other populations globally.

Individuals that contain both GSTM1*0 and GSTT1*0 genotypes, potentially have higher disfunctionality of these *GSTM1* and *GSTT1* genes than those having either one of the mutant genes. The data present in Table 6, column header 'Presence of both GSTM1*0 and GSTT1*0', shows individuals in the study that were homozygous for both mutations. This data along with that available in literature were used to compile the barchart illustration (Figure 20). With exception of JU and OVA, all study populations have individuals who are

homozygous for both mutations. The proportions of these individuals range from being low in Khoisan-speakers to slightly higher in Bantu-speakers of East and Central Africa. CAR_P (1.7%), too contain very low proportion of individuals who have this trait. Compared with African populations investigated in the study the number of individuals homozygous for both mutations is higher in SAW (11.1%). Such individuals appear to increase further north in Africa (Tunisia, 21%), being more frequent in the Middle East (Saudi Arabia, 22.7%; Lebanon, 16.3%; and Bahrain, 14.4%).

Table 6 GSTM1 and GSTT1 null genotype frequency

Population Code ¥	African region & country of sample	Presence of both GSTM1*0 & GSTT1*0	Presence of GSTM1*0	Presence of GSTT1*0	Total Sample
	Southern Africa				
SEB	South Africa	6 (5%)	25 (20.7%)	25 (20.7%)	121
XUN	South Africa & Nambia	2 (2.7%)	2 (2.7%)	25 (33.8%)	74
KHWE	South Africa & Nambia	2 (4.3%)	6 (13%)	19 (41.3%)	46
SAC	South Africa	5 (5.6%)	20 (22.2%)	17 (18.9%)	90
SAI	South Africa	6 (5.3%)	26 (23%)	12 (10.6%)	113
SAW	South Africa	5 (11.1%)	17 (37.8%)	4 (8.9%)	45
JU	Botswana	0 (0%)	0 (0%)	13 (39.4%)	33
NAMA	Namibia	1 (4.5%)	1 (4.5%)	9 (40.9%)	22
DAMA	Namibia	1 (2.7%)	2 (5.4%)	10 (27%)	37
HER	Namibia	4 (5.5%)	5 (6.8%)	29 (39.7%)	73
HIM	Namibia	2 (5.1%)	6 (15.4%)	16 (41%)	39
OVA	Namibia	0 (0%)	14 (22.6%)	17 (27.4%)	62
	Eastern Africa				
MOZ	Mozambique	4 (5.7%)	9 (12.9%)	16 (22.9%)	70
ZAM	Zambia	11 (6.2%)	29 (16.4%)	29 (16.4%)	177
UGA	Uganda	19 (8.1%)	40 (16.9%)	66 (28%)	236
	Central Africa				
CAR_U	Central African Republic	17 (8.8%)	32 (16.6%)	54 (28%)	193
CAR_P	Central African Republic	2 (1.7%)	22 (18.6%)	7 (5.9%)	118
DRC	Democratic Republic of Congo	16 (5.7%)	45 (15.9%)	68 (24%)	283

¥ For population name and ethnicity refer to table 1

NB: Frequecies indicate proportion of genotypes in populations

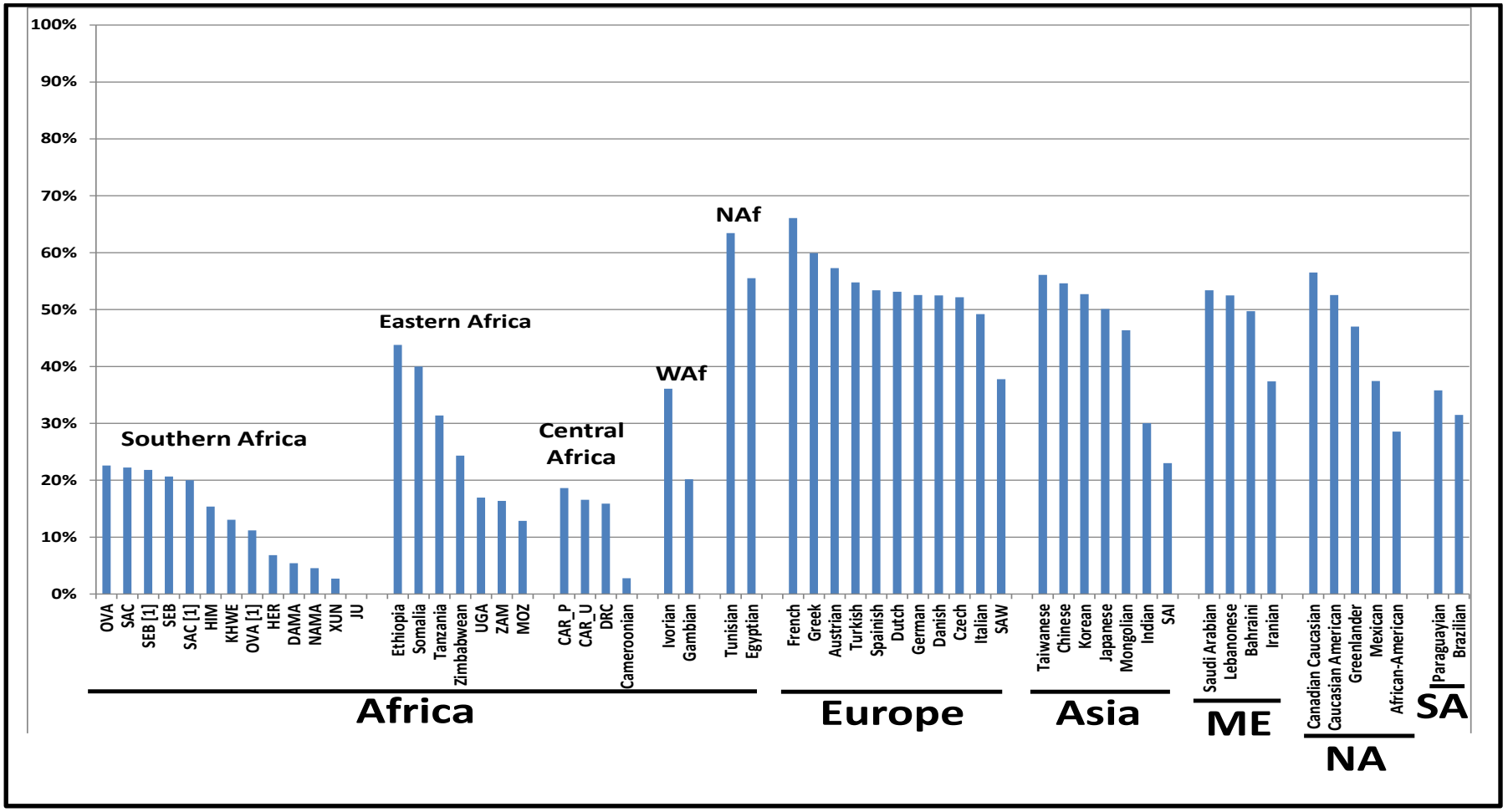

Figure 18 Bar chart showing GSTM1*0 genotype proportions worldwide
NB: WAf = Western Africa; NAf = Northern Africa; ME = Middle East; NA = North America; SA = South America

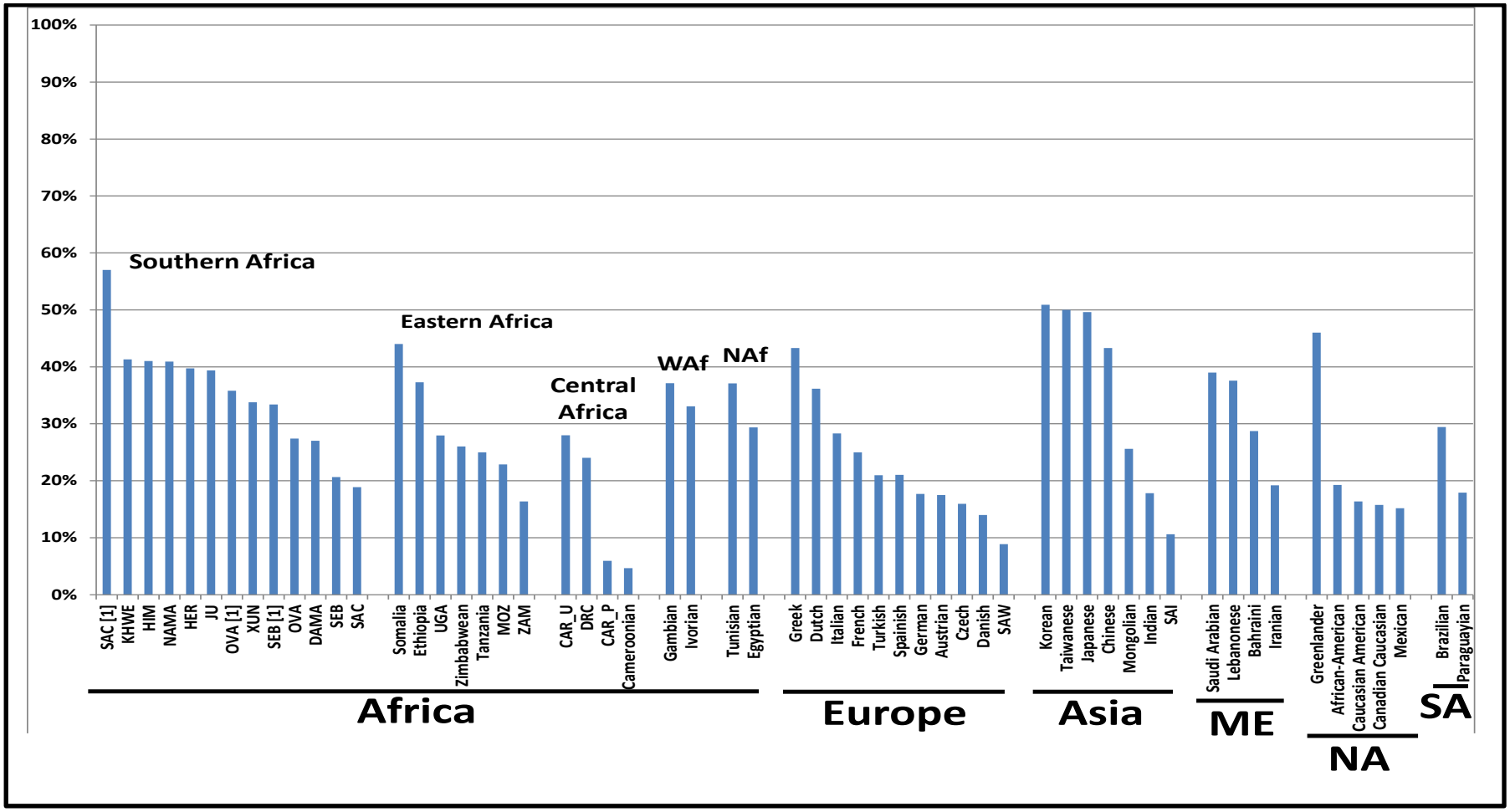

Figure 19 Bar chart showing GSTT1*0 genotype proportions worldwide
NB: WAf = Western Africa; NAf = Northern Africa; ME = Middle East; NA = North America; SA = South America

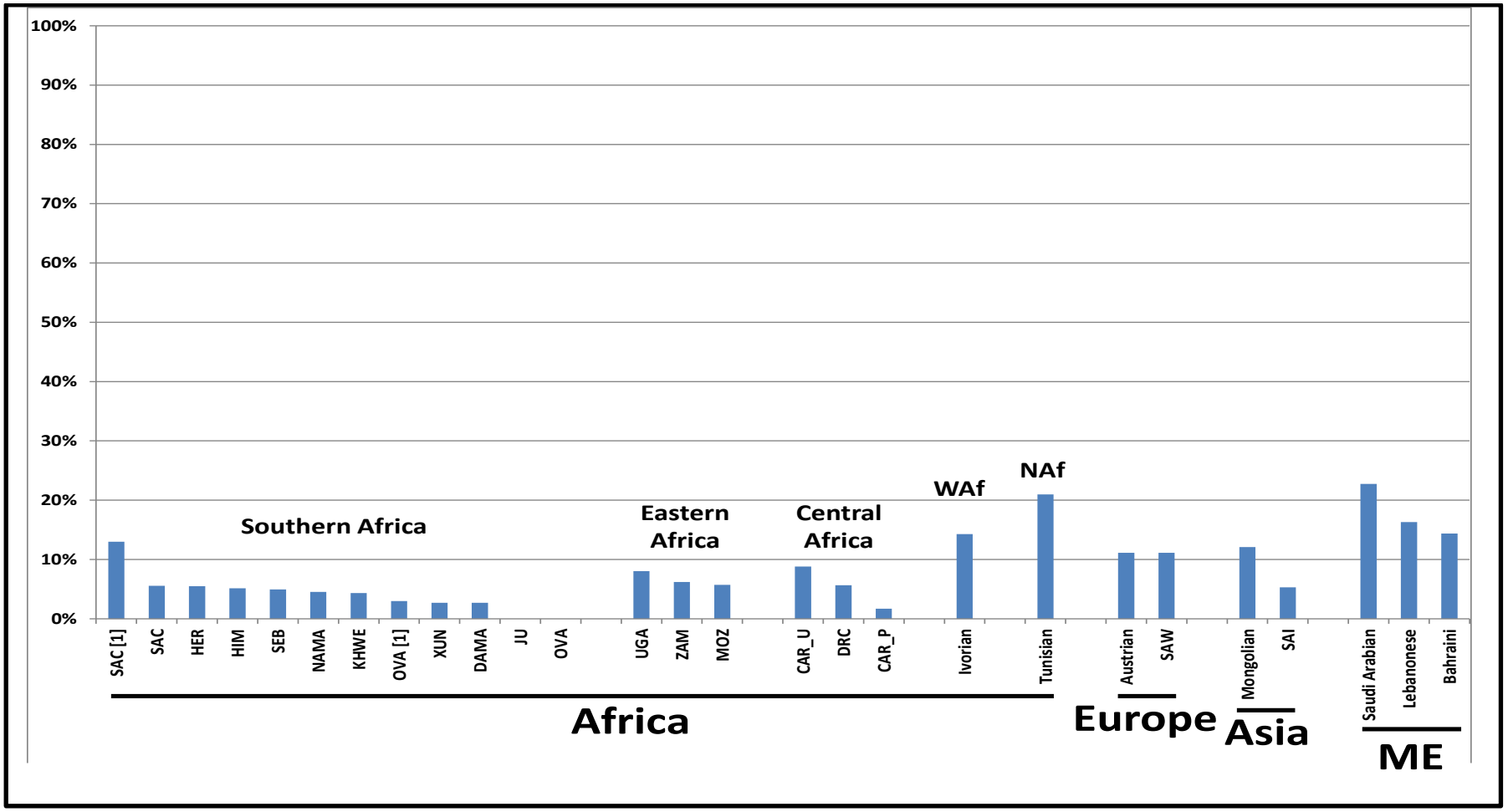

Figure 20 Bar chart showing the proportion of individuals who are homozygous for both GSTM1*0 and GSTT1*0 mutations worldwide
NB: WAf = Western Africa; NAf = Northern Africa; ME = Middle East; NA = North America; SA = South America

3.3.3. GSTP1*Ile105Val allele frequencies

This mutant allele is found at largest proportions in Khoe-San varying from 74.1% (JU) to 57.1% (KHWE) (Table 5 and Figure 21. Southern Bantu-speakers contain the allele at frequencies between 55.1% (HER) to 45.8% (OVA). Across Central, eastern and western Africa, GSTP1*Ile105Val is found in Bantu-speakers at frequencies ranging from 50% in Ghana to 16% in Tanzania (Table 5 and Figure 21 with comparative data in Appendix B, Table G). The allele is found in SAC at a frequency of 48.2%, slightly higher when compared to the South African Coloured (43%) group reported in Adams et al. (2003). This is probably due to the different histories for each group where the Coloured population was sampled primarily in Johannesburg and the other from Adams et al. (2003) in the Western Cape, South Africa. GSTP1*Ile105Val is found in SAI (27.6%) and SAW (38.6%) at similar similar proportions to those of Indians and Europeans, respectively.

Further north of Africa, the mutation decreases, where in Europe its proportion ranges from 36.3% (Netherlands) to 25.7% (Turkey) (Table 5 and Figure 21 with comparative data in Appendix B, Table G). In Asia the GSTP1*Ile105Val allele is found in Indians at a frequency of 27.3%, where its proportion decreases towards the far-East (China, 17.9; Taiwan, 17.8%; and Japan, 15.1%). In the Americas the allele appears to decrease from north to south, where it is present at frequencies from 48.1% in Mexico to 34.1% in Brazil. Caucasian Americans (34.1%) contain the allele at similar proportions as most European populations. Further, GSTP1*Ile105Val is found in African-Americans at a frequency of 42.2%, similar to that observed in African populations.

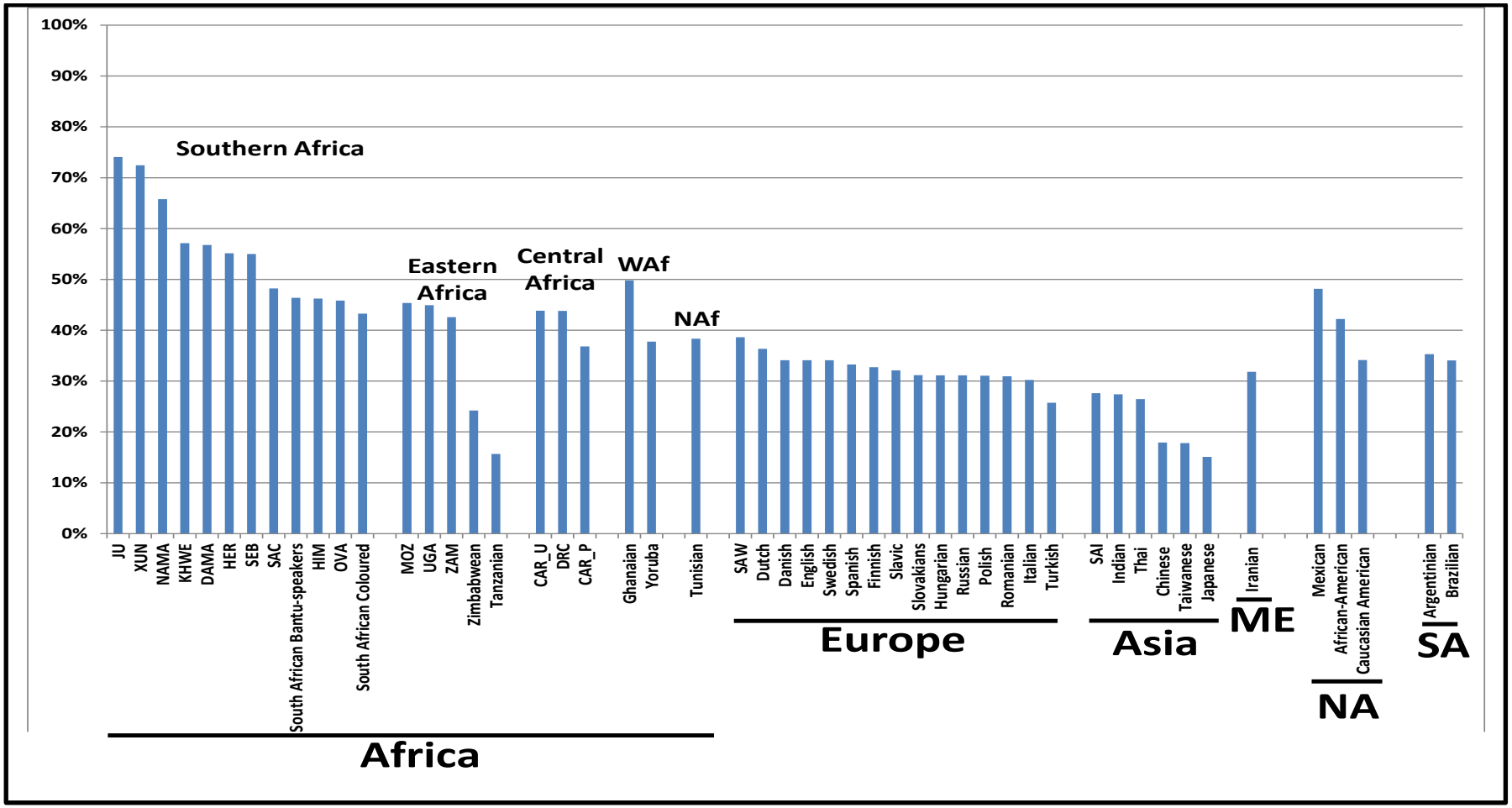

Figure 21 Bar chart showing GSTP1*Ile105Val allele proportions worldwide
NB: WAf = Western Africa; NAf = Northern Africa; ME = Middle East; NA = North America; SA = South America

3.3.4. GSTP1*Ala114Val allele frequencies

GSTP1*Ala114Val is primarily absent in Africans, with the exception of South African Bantu-speakers (13.4%; Li et al., 2010), SEB (1.4%), HER(1.3%), MOZ (1%), and UGA (1%) (Table 5 and Figure 22) . From the literature available the mutation is most frequent in Europeans ranging from 9.7% (Danish) to 5.1% (Spanish) (Table 5 and Figure 22 with comparative data in Appendix B, Table G). Similar to parental populations, the mutation is found at a frequency of 14.7% in SAW. The mutation is present in SAI (6.7%) and Indians (6.5%).

Its presence in Europeans and Indians, and near absence in Africa indicates the mutations was recently introduced by Europeans and/or Indians in SEB, HER, MOZ, and UGA populations. Similarly, this explains the mutation's presence in SAC at a frequency of 4.7% and the Coloured group (10.9%) from Li et al. (2010). African-Americans contain the mutation at a frequency of 2.7%. Given their ancestral history, and the absence in the African parental population it can be assumed that African-Americans inherited the mutation from their European descendants.

3.3.5. NAT2*14A allele frequencies

NAT2*14A is exclusively found in Africa with highest proportions in sub-Saharan Africans (Table 5 and Figure 23). The mutation is present in all Khoe-San (range 5.3% to 4.4%) populations except the JU. Among Bantu-speakers, the mutation is present at higher frequencies from 23% in Mbuti Pygmies of the DRC to 1.4% in DAMA. Towards the north of Africa the mutation tends to be low to absent (Somalia, 1%; Ethiopia, 0%; and Morocco, 0%).

Outside Africa, the trend seen in North Africa continues where the mutation is absent in Europe, Asia, the Middle East and the Americas (Table 5 and Figure 23

with comparative data in Appendix B, Table H). Exceptions to this trend are found in the Italians (1%), Slavics (0.1%), Spanish (0.1%), Cambodians (2%), African-Americans (8%), Nicaraguans (0.4%), Brazilians (4%) and Colombians (2%). Africans were the likely contributors of this allele to these populations. Further, argument of African contribution can be made for South African Coloured populations, given their history being born from European, Asian and African descendants. This distribution pattern observed in this study for the NAT2*14A allele has been reported before (Delomenie et al., 1996; Bayoumi et al., 1997; Butcher et al., 2002; Adams et al., 2003; Cavaco et al., 2007; Yoshida et al., 2007; Sabbagh et al., 2008; Dandara et al., 2010; Staehli et al., 2013; Touré et al., 2012).

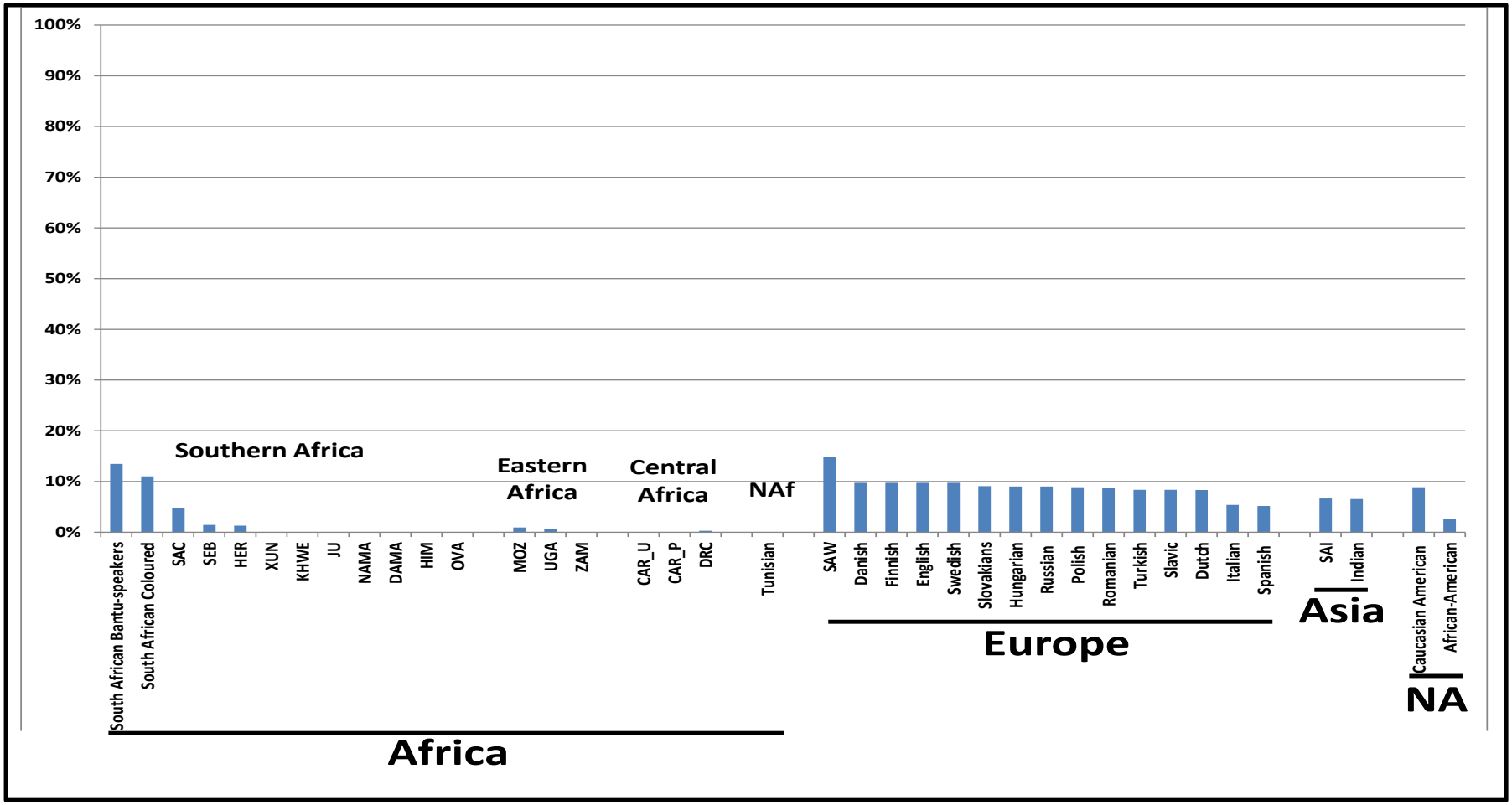

Figure 22 Bar chart showing GSTP1*Ala114Val allele proportions worldwide
NB: NAf = Northern Africa; NA = North America

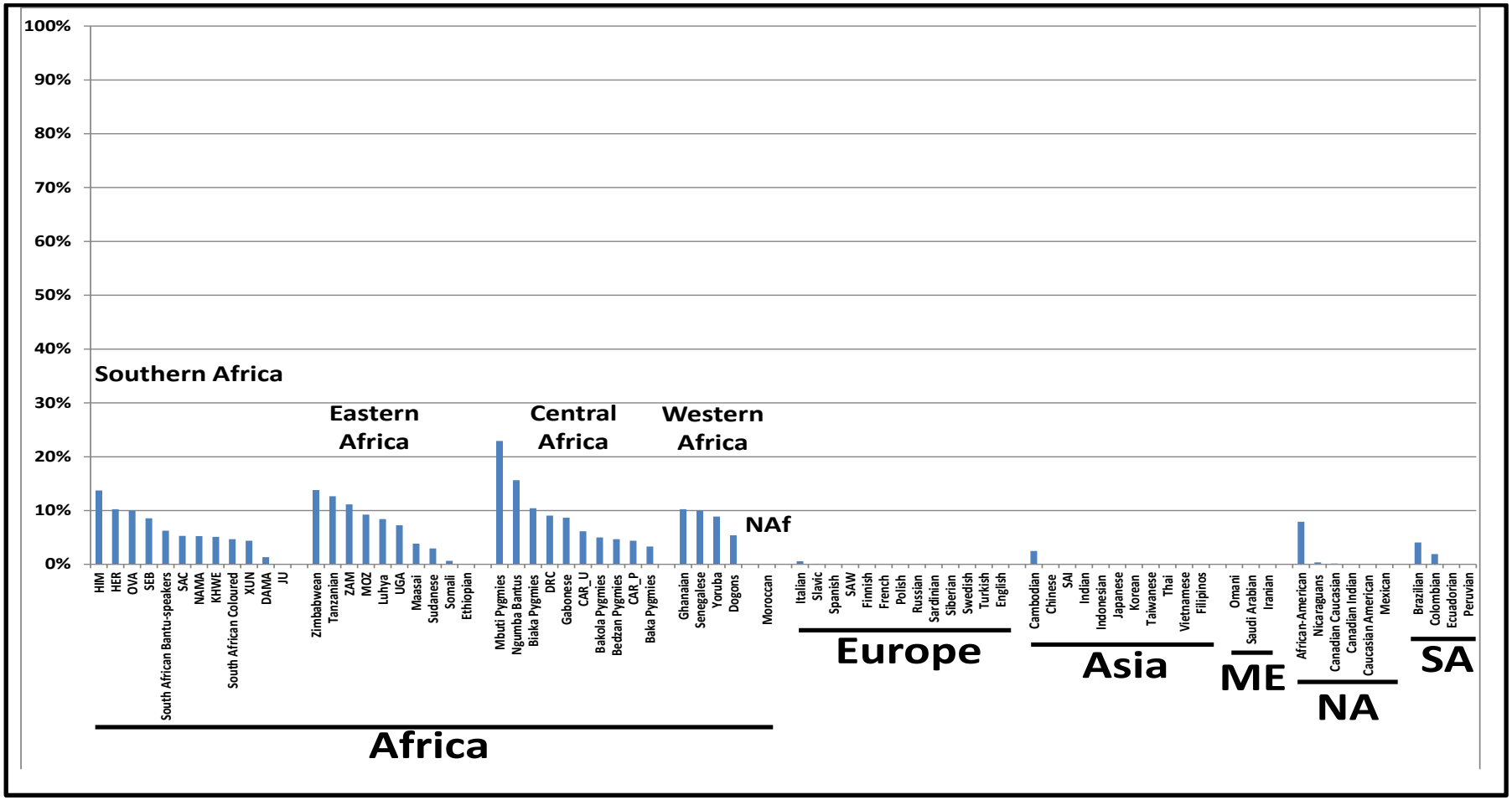

Figure 23 Bar chart showing NAT2*14A allele proportions worldwide
NB: NAf = Northern Africa; ME = Middle East; NA = North America; SA = South America

3.4. Hardy-Weinberg Equilibrium and Tajima's D statistic tests performed

The allele frequencies for the 13 loci were evaluated for Hardy-Weinberg Equilibrium (HWE) and Tajima's D statistic with a 95% confidence interval, for each of the 18 populations used in this study. The results of these tests are given in the Table 7 to Table 24, where significant tests p-values have been highlighted in blue. Loci CYP2A6*8, CYP2D6*3A, CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I) were all monomorphic containing only wild-type allele for all 18 populations, and therefore could not be examined for HWE. In the SEB, the GSTP1*Ala114Val was the only locus found to deviate from HWE (Table 11). HER population was found to deviate from HWE for the GSTP1*Ile105Val locus (Table 13). The ZAM populations deviated from HWE proportions for the NAT2*14A locus (Table 18). The CYP1A1*2A marker was found to deviate from HWE proportions in both CAR_U and DRC populations (Table 19 and Table 20). In addition, the DRC population deviated from HWE for loci CYP1A2*1C and CYP2A6*7 (Table 19). The SAI were the only sea-borne population that showed to have deviated from HWE, this for the CYP1A2*1F marker (Table 24). Due to the low amplification success for CYP1A2*1C, the JU population contained only two alleles for this locus, both were mutant alleles and therefore HWE test was not conducted (Table 7).

As listed above, there where some loci that were found to deviate from HWE in the various populations. However, when allelic data for the 18 populations were applied to Tajima's D statistic test all loci except one was not significant for selective neutrality. CYP1A1*2C was found to be under selection within the XUN population (Table 8), D=-0.9921 (p=0.0000). The negative D-value is an indication for this locus there is an excess of lower frequency alleles and the population size is at expansion probably following some bottleneck event. This is surprising as the XUN contains only one mutant allele (1%) among the 69 individuals typed for

this marker. Given that the marker was not found to deviate from HWE (Table 8), it is possible that the mutation is not under selection.

Table 7 HWE and Tajima's D statistic tests performed for JU population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	**Sample size (2n)**	**Observed Heterozygosity**	**Expected Heterozygosity**	**P-value**	**Tajima's D statistic**	**P-value**
CYP1A1*2A	52	0.6154	0.4344	0.0581	1.2807	0.8677
CYP1A1*2C	54	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	2	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1F	54	0.5926	0.4249	0.0586	1.2332	0.8609
CYP2A6*7	54	0.2222	0.2013	1.0000	-0.1092	0.2789
CYP2A6*8	54	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	44	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	54	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Pst* I)	52	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	30	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	54	0.4444	0.3913	0.6395	1.0318	0.8371
GSTP1*Ala114Val	54	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	54	Monomorphic - test not done			0.0000	1.0000

Note 2n indicates the total number of alleles used in tests

Table 8 HWE and Tajima's D statistic tests performed for XUN population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	**Sample size (2n)**	**Observed Heterozygosity**	**Expected Heterozygosity**	**P-value**	**Tajima's D statistic**	**P-value**
CYP1A1*2A	134	0.5522	0.4818	0.3102	1.7731	0.9423
CYP1A1*2C	138	0.0145	0.0145	1.0000	-0.9921	0.0000
CYP1A2*1C	58	0.4138	0.4356	1.0000	1.3151	0.8724
CYP1A2*1F	138	0.5217	0.4477	0.1908	1.5769	0.8960
CYP2A6*7	138	0.1594	0.1478	1.0000	-0.2017	0.2690
CYP2A6*8	138	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	122	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	138	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Pst* I)	132	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	122	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	138	0.4348	0.4020	0.5579	1.3059	0.8583
GSTP1*Ala114Val	138	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	136	0.0588	0.0850	0.1109	-0.5771	0.2131

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Table 9 HWE and Tajima's D statistic tests performed for KHWE population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	98	0.4490	0.4124	0.7267	1.2971	0.8668
CYP1A1*2C	98	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	60	0.3667	0.4401	0.4112	1.3506	0.8751
CYP1A2*1F	98	0.6122	0.4797	0.0709	1.6968	0.9335
CYP2A6*7	98	0.0612	0.0600	1.0000	-0.7953	0.1557
CYP2A6*8	98	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	94	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	98	0.0612	0.0600	1.0000	-0.7953	0.1626
CYP2E1*5B(*Pst* I)	98	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	88	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	98	0.4082	0.4949	0.2487	1.7867	0.9597
GSTP1*Ala114Val	98	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	98	0.1020	0.0978	1.0000	-0.5704	0.2147

Note 2n indicates the total number of alleles used in tests

Table 10 HWE and Tajima's D statistic tests performed for NAMA population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	38	0.1053	0.1935	0.1627	-0.2712	0.2442
CYP1A1*2C	38	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	18	0.1111	0.1111	1.0000	-1.1647	0.1299
CYP1A2*1F	38	0.4737	0.4225	1.0000	1.1231	0.8484
CYP2A6*7	38	Monomorphic - test not done			0.0000	1.0000
CYP2A6*8	38	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	36	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	38	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Pst* I)	38	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	34	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	38	0.3684	0.4623	0.6059	1.3656	0.8766
GSTP1*Ala114Val	38	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	38	0.1053	0.1024	1.0000	-0.8255	0.1631

Note 2n indicates the total number of alleles used in tests

Table 11 HWE and Tajima's D statistic tests performed for SEB population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	**Sample size (2n)**	**Observed Heterozygosity**	**Expected Heterozygosity**	**P-value**	**Tajima's D statistic**	**P-value**
CYP1A1*2A	140	0.3571	0.3473	1.0000	0.9842	0.8208
CYP1A1*2C	140	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	66	0.4242	0.4028	1.0000	1.1506	0.8475
CYP1A2*1F	140	0.4429	0.4862	0.4668	1.8083	0.9449
CYP2A6*7	140	0.0143	0.0143	1.0000	-0.9906	0.0725
CYP2A6*8	140	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	134	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	140	0.0286	0.0284	1.0000	-0.9071	0.1195
CYP2E1*5B(*Pst* I)	138	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	136	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	140	0.4714	0.4986	0.8083	1.8814	0.9690
GSTP1*Ala114Val	140	0.0000	0.0284	0.0063	-0.9071	0.1113
NAT2*14A	140	0.1143	0.1579	0.0680	-0.1391	0.2952

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Table 12 HWE and Tajima's D statistic tests performed for MOZ population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	**Sample size (2n)**	**Observed Heterozygosity**	**Expected Heterozygosity**	**P-value**	**Tajima's D statistic**	**P-value**
CYP1A1*2A	104	0.3269	0.3254	1.0000	0.7937	0.7965
CYP1A1*2C	108	0.0185	0.0185	1.0000	-1.0193	0.0746
CYP1A2*1C	80	0.1500	0.1823	0.3201	-0.1168	0.2862
CYP1A2*1F	108	0.4444	0.5047	0.4261	1.8653	1.0000
CYP2A6*7	108	0.0556	0.0545	1.0000	-0.8057	0.1550
CYP2A6*8	108	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	106	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	108	0.0556	0.0545	1.0000	-0.8057	0.1527
CYP2E1*5B(*Pst* I)	104	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	108	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	108	0.5370	0.5004	0.7814	1.8396	0.9757
GSTP1*Ala114Val	108	0.0185	0.0185	1.0000	-1.0193	0.0734
NAT2*14A	108	0.1482	0.1696	0.3670	-0.1228	0.2841

Note 2n indicates the total number of alleles used in tests

Table 13 HWE and Tajima's D statistic tests performed for HER population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	156	0.3846	0.3838	1.0000	1.2218	0.8521
CYP1A1*2C	156	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	78	0.3846	0.4732	0.3048	1.6088	0.9160
CYP1A2*1F	156	0.5897	0.4764	0.0539	1.7714	0.9322
CYP2A6*7	156	0.0128	0.0128	1.0000	-0.9787	0.0761
CYP2A6*8	156	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	154	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	156	0.0641	0.0625	1.0000	-0.6844	0.1936
CYP2E1*5B(*Pst* I)	152	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	154	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	156	0.6154	0.4979	0.0395	1.8989	0.9733
GSTP1*Ala114Val	156	0.0256	0.0255	1.0000	-0.9037	0.1094
NAT2*14A	156	0.1539	0.1853	0.1703	0.0443	0.7032

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Table 14 HWE and Tajima's D statistic tests performed for HIM population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	80	0.4000	0.3532	0.6523	0.9001	0.8147
CYP1A1*2C	80	0.0250	0.0250	1.0000	-1.0527	0.0823
CYP1A2*1C	72	0.5000	0.4507	0.7080	1.4568	0.8959
CYP1A2*1F	80	0.5500	0.5063	0.7495	1.8116	1.0000
CYP2A6*7	80	0.0250	0.0250	1.0000	-1.0527	0.0813
CYP2A6*8	80	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	78	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	80	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Pst* I)	80	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	78	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	80	0.6250	0.5035	0.2026	1.7946	0.9814
GSTP1*Ala114Val	80	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	80	0.2750	0.2402	1.0000	0.2279	0.7326

Note 2n indicates the total number of alleles used in tests

Table 15 HWE and Tajima's D statistic tests performed for OVA population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	120	0.3667	0.3423	0.7157	0.9238	0.8127
CYP1A1*2C	120	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	46	0.1739	0.2937	0.0977	0.3995	0.7588
CYP1A2*1F	120	0.5167	0.5029	1.0000	1.8765	0.9894
CYP2A6*7	120	Monomorphic - test not done			0.0000	1.0000
CYP2A6*8	120	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	120	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	120	0.0500	0.0492	1.0000	-0.8149	0.1200
CYP2E1*5B(*Pst* I)	120	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	120	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	120	0.3833	0.5007	0.0749	1.8632	0.9765
GSTP1*Ala114Val	120	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	120	0.2000	0.1815	1.0000	-0.0299	0.2999

Note 2n indicates the total number of alleles used in tests

Table 16 HWE and Tajima's D statistic tests performed for DAMA population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	74	0.3514	0.3588	1.0000	0.9152	0.8213
CYP1A1*2C	74	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	68	0.2941	0.2950	1.0000	0.5144	0.7685
CYP1A2*1F	74	0.5135	0.4832	0.7426	1.6564	0.9315
CYP2A6*7	74	Monomorphic - test not done			0.0000	1.0000
CYP2A6*8	74	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	74	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	74	0.0541	0.0533	1.0000	-0.9046	0.1319
CYP2E1*5B(*Pst* I)	74	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	68	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	74	0.5405	0.4976	0.7386	1.7424	0.9580
GSTP1*Ala114Val	74	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	74	0.0270	0.0270	1.0000	-1.0613	0.0831

Note 2n indicates the total number of alleles used in tests

Table 17 HWE and Tajima's D statistic tests performed for UGA population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	448	0.2277	0.2358	0.5712	0.5171	0.7793
CYP1A1*2C	454	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	320	0.2500	0.2977	0.0609	0.8369	0.8084
CYP1A2*1F	454	0.5066	0.4992	0.8924	2.1019	0.9854
CYP2A6*7	454	0.0661	0.0640	1.0000	-0.5128	0.2414
CYP2A6*8	454	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	452	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	454	0.0617	0.0599	1.0000	-0.5376	0.2380
CYP2E1*5B(*Pst* I)	450	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	446	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	454	0.5374	0.4960	0.2246	2.0824	0.9715
GSTP1*Ala114Val	454	0.0132	0.0132	1.0000	-0.8185	0.1140
NAT2*14A	454	0.1454	0.1351	0.6151	-0.0858	0.3211

Note 2n indicates the total number of alleles used in tests

Table 18 HWE and Tajima's D statistic tests performed for ZAM population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	350	0.3714	0.3913	0.5648	1.4111	0.8715
CYP1A1*2C	350	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	274	0.2920	0.3350	0.1946	1.0334	0.8291
CYP1A2*1F	350	0.5543	0.4972	0.1774	2.0446	0.9792
CYP2A6*7	350	0.0343	0.0338	1.0000	-0.7276	0.1650
CYP2A6*8	350	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	344	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	350	0.0743	0.0717	1.0000	-0.5007	0.2413
CYP2E1*5B(*Pst* I)	350	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	340	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	350	0.5200	0.4904	0.4378	2.0035	0.9670
GSTP1*Ala114Val	350	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	350	0.1543	0.1986	0.0090	0.2582	0.7488

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Table 19 HWE and Tajima's D statistic tests performed for DRC population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	320	0.2625	0.3429	0.0053	1.1065	0.8309
CYP1A1*2C	342	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	66	0.2424	0.4028	0.0305	1.1506	0.8491
CYP1A2*1F	340	0.5000	0.4985	1.0000	2.0472	0.9823
CYP2A6*7	342	0.4737	0.3626	0.0000	1.2351	0.8459
CYP2A6*8	342	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	258	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	342	0.1404	0.1309	1.0000	-0.1501	0.3074
CYP2E1*5B(*Pst* I)	242	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	246	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	338	0.5089	0.4937	0.7544	2.0175	0.9660
GSTP1*Ala114Val	342	0.0059	0.0059	1.0000	-0.8978	0.0616
NAT2*14A	342	0.1462	0.1653	0.1441	0.0559	0.7173

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Table 20 HWE and Tajima's D statistic tests performed for CAR_U population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	354	0.2881	0.3477	0.0277	1.1520	0.8419
CYP1A1*2C	358	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	174	0.2414	0.2794	0.2432	0.6230	0.7832
CYP1A2*1F	358	0.5028	0.4931	0.8780	2.0239	0.9674
CYP2A6*7	358	0.2346	0.2077	0.1400	0.3160	0.7536
CYP2A6*8	358	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	328	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	358	0.0615	0.0597	1.0000	-0.5694	0.2180
CYP2E1*5B(*Pst* I)	342	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	330	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	358	0.5084	0.4938	0.7694	2.0282	0.9676
GSTP1*Ala114Val	358	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	358	0.1229	0.1157	1.0000	-0.2346	0.2882

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Table 21 HWE and Tajima's D statistic tests performed for CAR_P population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	162	0.3210	0.3865	0.1423	1.2450	0.8560
CYP1A1*2C	182	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1C	50	0.6000	0.4580	0.1783	1.4124	0.8926
CYP1A2*1F	182	0.4286	0.3629	0.1387	1.1269	0.8395
CYP2A6*7	182	0.0769	0.0744	1.0000	-0.5858	0.2080
CYP2A6*8	182	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	180	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	182	0.0659	0.0641	1.0000	-0.6468	0.1965
CYP2E1*5B(*Pst* I)	164	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	144	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	182	0.5165	0.4678	0.3742	1.7497	0.9248
GSTP1*Ala114Val	182	Monomorphic - test not done			0.0000	1.0000
NAT2*14A	182	0.0659	0.0845	0.1489	-0.5257	0.2161

Note 2n indicates the total number of alleles used in tests

Table 22 HWE and Tajima's D statistic tests performed for SAC population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	170	0.3177	0.3289	0.7454	0.9127	0.8174
CYP1A1*2C	170	0.1059	0.1009	1.0000	-0.4407	0.2380
CYP1A2*1C	84	0.0714	0.0697	1.0000	-0.7743	0.1621
CYP1A2*1F	170	0.3882	0.4317	0.4469	1.5224	0.8865
CYP2A6*7	170	0.1177	0.1114	1.0000	-0.3783	0.2569
CYP2A6*8	170	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	156	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	170	0.1059	0.1009	1.0000	-0.4407	0.2449
CYP2E1*5B(*Pst* I)	166	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	150	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	170	0.5177	0.5023	0.8283	1.9417	0.9902
GSTP1*Ala114Val	170	0.0941	0.0902	1.0000	-0.5039	0.2325
NAT2*14A	170	0.0824	0.1009	0.1986	-0.4407	0.2411

Note 2n indicates the total number of alleles used in tests

Table 23 HWE and Tajima's D statistic tests performed for SAW population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	84	0.2381	0.2479	1.0000	0.2850	0.7355
CYP1A1*2C	88	0.1364	0.1672	0.2956	-0.1835	0.2681
CYP1A2*1C	20	Monomorphic - test not done			0.0000	1.0000
CYP1A2*1F	88	0.4318	0.4114	1.0000	1.2682	0.8509
CYP2A6*7	88	0.0682	0.0666	1.0000	-0.7813	0.1704
CYP2A6*8	88	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	68	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	86	0.2093	0.1896	1.0000	-0.0557	0.2737
CYP2E1*5B(*Pst* I)	74	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	68	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	88	0.5000	0.4796	1.0000	1.6734	0.9331
GSTP1*Ala114Val	88	0.2500	0.2547	1.0000	0.3366	0.7352
NAT2*14A	88	Monomorphic - test not done			0.0000	1.0000

Note 2n indicates the total number of alleles used in tests

Table 24 HWE and Tajima's D statistic tests performed for SAI population

		Hardy-Weinberg Equilibrium test			Tajima's D test for neutral selection	
Locus	Sample size (2n)	Observed Heterozygosity	Expected Heterozygosity	P-value	Tajima's D statistic	P-value
CYP1A1*2A	206	0.4078	0.4444	0.5019	1.6341	0.8993
CYP1A1*2C	210	0.2191	0.2392	0.4105	0.4180	0.7638
CYP1A2*1C	116	0.0690	0.0672	1.0000	-0.7152	0.1778
CYP1A2*1F	210	0.3238	0.4692	0.0018	1.7848	0.9279
CYP2A6*7	210	0.0952	0.1083	0.2823	-0.3601	0.2638
CYP2A6*8	210	Monomorphic - test not done			0.0000	1.0000
CYP2D6*3	206	Monomorphic - test not done			0.0000	1.0000
CYP2D6*4	208	0.1058	0.1007	1.0000	-0.4069	0.2477
CYP2E1*5B(*Pst* I)	206	Monomorphic - test not done			0.0000	1.0000
CYP2E1*5B(*Rsa* I)	150	Monomorphic - test not done			0.0000	1.0000
GSTP1*Ile105Val	210	0.4571	0.4017	0.2234	1.3840	0.8731
GSTP1*Ala114Val	210	0.1333	0.1250	1.0000	-0.2605	0.2793
NAT2*14A	210	Monomorphic - test not done			0.0000	1.0000

Note 2n indicates the total number of alleles used in tests

Significant tests p-values have been highlighted in blue

Given that majority of the sample had not deviated from HWE proportions, we used the exact test (Raymond and Rousset, 1995), reported to be more accurate when expected outcomes are small. Table 25 shows that most (58.8%) pair-wise comparisons are significant ($p<0.05$, yellow highlighted combinations). These significant combinations indicate these populations are significantly different to each other, possibly due to differing population sizes and/or allelic frequency differences.

Table 25 Exact Test of sample differentiation based on allele frequencies

	CAR_U	CAR_P	DAMA	DRC	HER	HIM	JU	KHWE	MOZ	NAMA	OVA	SAC	SAI	SAW	SEB	UGA	XUN
CAR_P	0.01270+0.0060																
DAMA	0.83760+0.0581	0.00000+0.0000															
DRC	0.00000+0.0000	0.01800+0.0127	0.00000+0.0000														
HER	0.34780+0.0513	0.00910+0.0055	0.22155+0.0241	0.06310+0.0451													
HIM	0.40820+0.0783	0.00000+0.0000	0.20485+0.0217	0.00000+0.0000	0.70520+0.0384												
JU	0.02270+0.0213	0.06180+0.0209	0.00000+0.0000	0.35065+0.1045	0.00890+0.0043	0.00000+0.0000											
KHWE	0.24705+0.0620	0.05905+0.0245	0.00150+0.0010	0.05535+0.0360	0.59140+0.0471	0.54765+0.0349	0.03955+0.0148										
MOZ	0.66565+0.0832	0.00000+0.0000	0.57320+0.0257	0.00040+0.0004	0.04710+0.0231	0.00820+0.0075	0.00000+0.0000	0.00000+0.0000									
NAMA	0.34905+0.0508	0.74520+0.0312	0.00235+0.0024	0.56835+0.0921	0.42695+0.0382	0.07505+0.0152	0.22330+0.0268	0.75710+0.0326	0.01770+0.0087								
OVA	0.74320+0.0461	0.13260+0.0223	0.03515+0.0088	0.10350+0.0356	0.65460+0.0346	0.01730+0.0048	0.00405+0.0025	0.53490+0.0238	0.09570+0.0233	0.27895+0.0387							
SAC	0.10955+0.0287	0.18950+0.0476	0.00000+0.0000	0.00485+0.0017	0.02715+0.0108	0.00160+0.0012	0.00005+0.0001	0.18985+0.0265	0.00130+0.0011	0.91600+0.0233	0.33100+0.0268						
SAI	0.00225+0.0025	0.00000+0.0000	0.19880+0.0591	0.00000+0.0000	0.00000+0.0000	0.00705+0.0062	0.00825+0.0047	0.00865+0.0029	0.12530+0.0320	0.14980+0.0316	0.02885+0.0100	0.16195+0.0423					
SAW	0.00000+0.0000	0.08060+0.0323	0.00040+0.0004	0.25110+0.0776	0.00600+0.0055	0.02155+0.0082	0.02760+0.0064	0.02650+0.0188	0.00000+0.0000	0.66825+0.0498	0.02850+0.0120	0.41190+0.0411	0.03470+0.0161				
SEB	0.42270+0.0610	0.21870+0.0358	0.04120+0.0086	0.05415+0.0191	0.63265+0.0685	0.09230+0.0300	0.00240+0.0022	0.24155+0.0301	0.08835+0.0319	0.81170+0.0312	0.14085+0.0253	0.53380+0.0486	0.00000+0.0000	0.00525+0.0050			
UGA	0.05005+0.0287	0.00000+0.0000	0.67750+0.0475	0.00000+0.0000	0.01720+0.0106	0.05415+0.0189	0.00000+0.0000	0.00000+0.0000	0.14610+0.0419	0.07210+0.0286	0.02385+0.0116	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000		
XUN	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000	0.04375+0.0323	0.01205+0.0046	0.26385+0.0243	0.11575+0.0273	0.00000+0.0000	0.11795+0.0410	0.00695+0.0073	0.00225+0.0017	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000	
ZAM	0.55670+0.0512	0.00000+0.0000	0.94350+0.0210	0.00000+0.0000	0.04205+0.0206	0.29400+0.0717	0.00000+0.0000	0.00000+0.0000	0.17000+0.0548	0.12235+0.0406	0.02880+0.0207	0.00000+0.0000	0.00000+0.0000	0.00000+0.0000	0.02445+0.0182	0.00315+0.0018	0.00000+0.0000

NB: For each combination, the first value is the p-value and the value following the '±' is the standard deviation. Yellow highlighted combinations represent significant combinations between two population groups ($p<0.05$)

3.5. Average gene diversity and average heterozygosity

The average gene diversity calculated across the 13 SNP loci for each of the study populations (Figure 24). In Figure 24, above each bar the average diversity is given, with the bracket line above indicating the standard deviation. Average diversity ranges between 0.2049±0.1340 in DRC to 0.1170±0.0864 in the CAR_P. San populations have intermediate average diversity indices within the sample, ranging between 0.1549±0.1031 in XUN to 0.1420±0.0977 in the JU. The average gene diversity in sea-borne migrants SAW (0.1985±0.1283) and SAI (0.1700±0.1081) are the second and fifth highest, respectively in the sample. This is possibly due to these populations having larger proportions of the less frequent allele for CYP1A1*2C, CYP2D6*4, GSTP1*Ile105Val and GSTP1*Ala114Val loci compared to indigenous African populations in the study (Table 5). Overall, all populations show fairly low gene diversity across the loci investigated with large standard deviations. This can be due to CYP2A6*8, CYP2D6*3A, CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I) loci monomorphic nature, accounting for 30.8% of the loci investigated in the study.

The average observed heterozygosity within each population is given in Figure 25, calculated from the observed heterozygosity indices for each of the 13 loci (Table 7 to Table 24). This is a measure of genetic variability, where the JU have the highest (0.4687±1809). Other Khoe-San populations and the CAR_P have similar average observed heterozygosity indices to that found in neighbouring Bantu-speaking populations. The lowest genetic variability is found at 0.1950±1943 in the MOZ. All populations had large standard deviations possibly due to many monomorphic loci being present; this more so for the JU and NAMA where both populations had only 4 polymorphic loci. In the sample an average of 6 polymorphic loci was observed. Sea-borne migrants SAW and SAI both had 7 polymorphic loci (Table 23 and Table 24), and the admixed SAC have the most (8 polymorphic loci, Table 22).

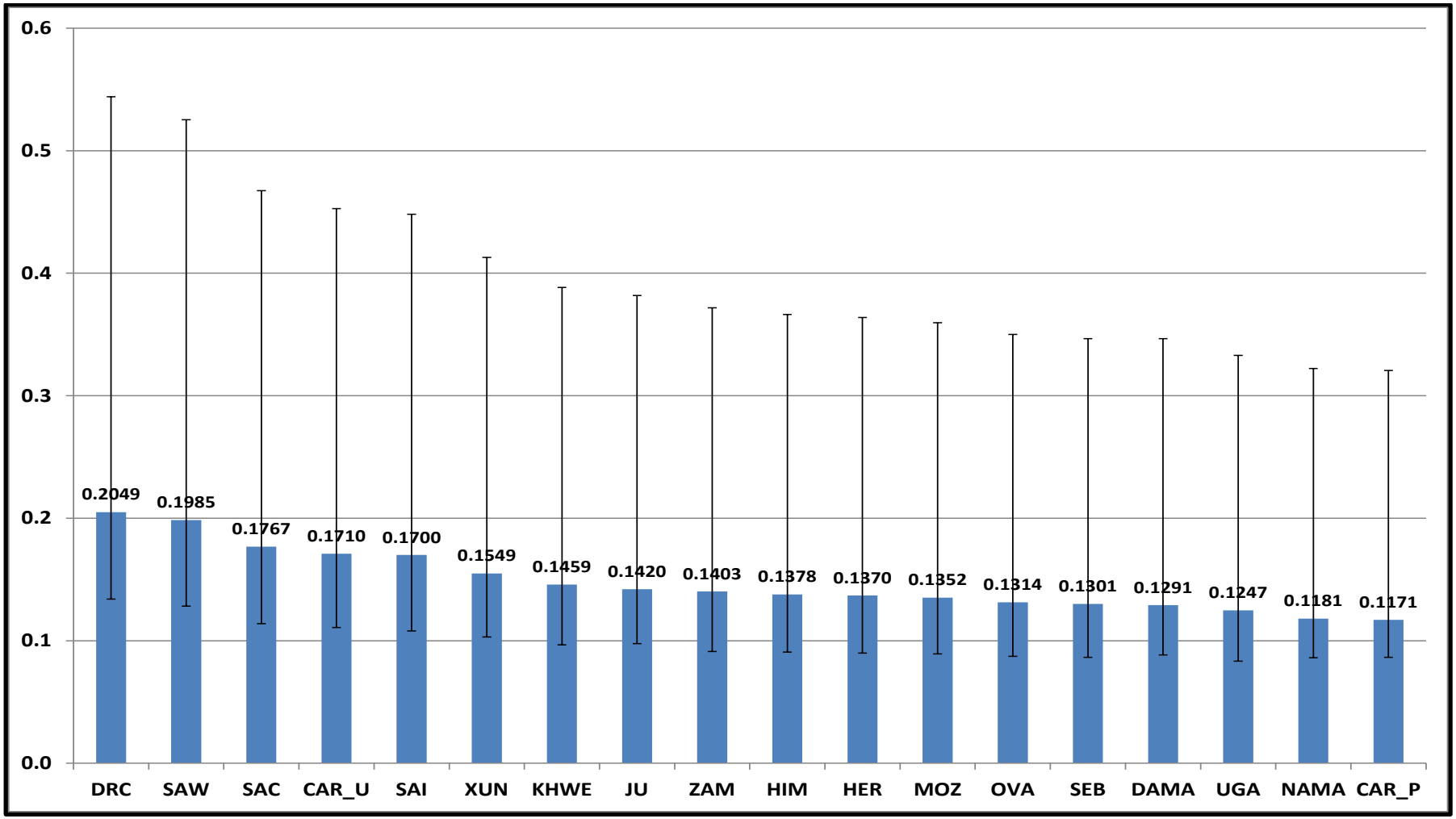

Figure 24 Average gene diversity across 13 SNP loci in study populations

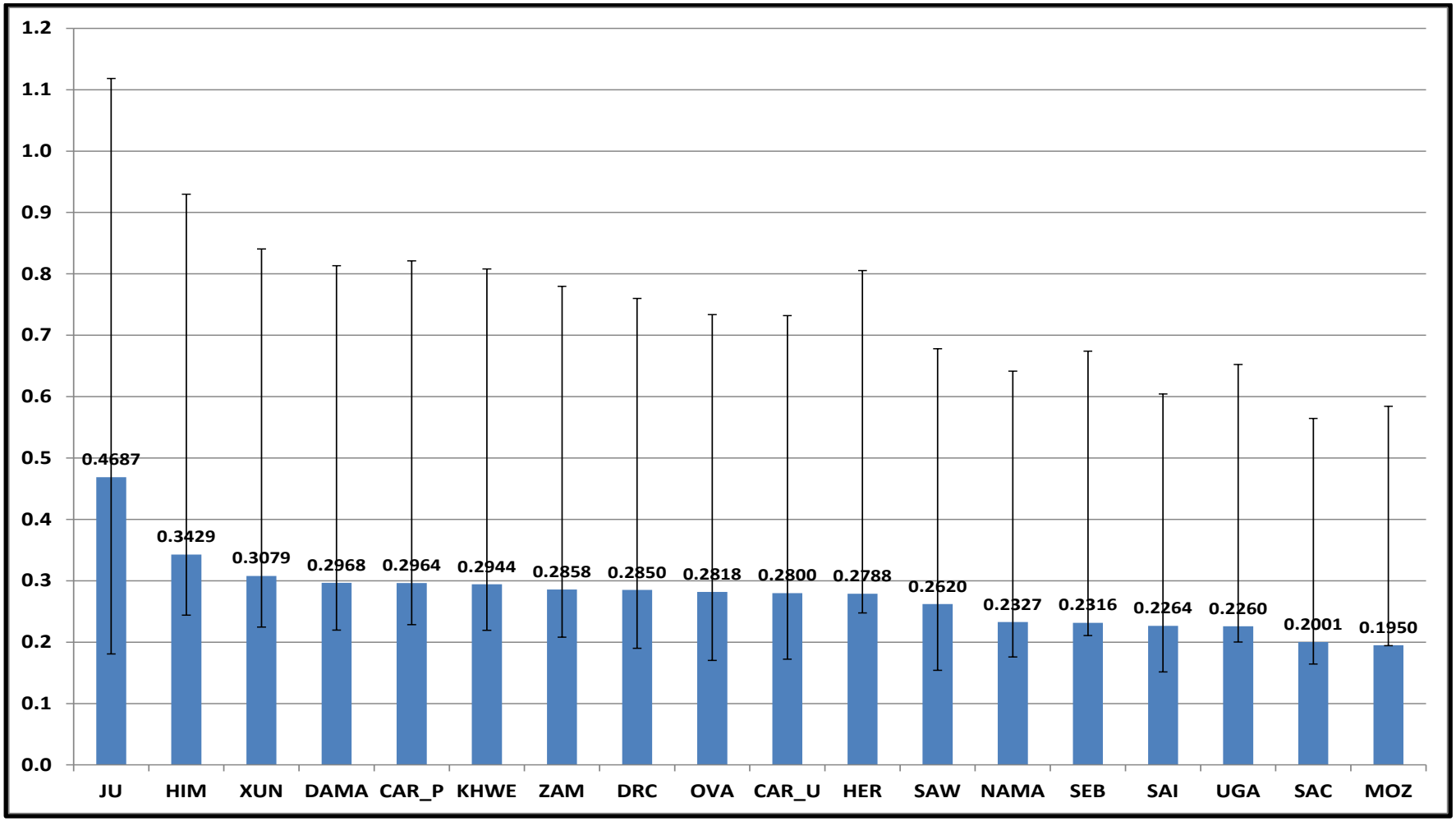

Figure 25 Average observed heterozygosity across 13 SNP loci in study population

3.6. Population affinities and evolution measured by genetic distance

To understand the genetic variation that exists among sub-Saharan Africans in the study, allele frequencies for the 13 SNP loci were used to determine genetic distances among indigenous African populations in this study (excluding sea-borne migrants SAI and SAW). Pair-wise F_{st} genetic distances were calculated (Appendix D) and used to understand the genetic affinities of the 16 indigenous African populations.

For the evolutionary relationships of indigenous African populations in the study, the genetic distance matrix (Appendix D) was used to compute a Neighbour-Joining population tree. Figure 26 gives the result of this analysis, showing the minimum evolutionary relationship of each population to each other as measured from genetic distances across the 13 SNP loci. Evident from the population tree the branch lengths are largest for the Khoe-San (except KHWE), where JU and XUN appear to be the oldest populations. CAR_P have a shorter branch length than Khoe-San, but much greater than other populations. The placement of branches of JU, XUN, NAMA and CAR_P indicates that these populations are the oldest and share a closer evolutionary history to each other than to Bantu-speakers. An exception to the Khoe-San can be seen for the KHWE who are closer to HER population, consistent with previous reports (Schlebusch, 2010a). Bantu-speakers are closely related and are much younger than Khoe-San and Pygmy populations. Similar to the associations seen in the PCA plots, the DAMA are more closely related to Bantu-speakers than to their linguistically related NAMA. In relation to the branch topology in the tree there is no concordance with the geographic location among the Bantu-speakers. The SAC branch is closely placed to SEB, indicating close evolutionary history with the SAC slightly older than SEB. This assessment though as in the PCA plots, needs to be

treated with caution as the sea-borne parental populations SAW and SAI have been excluded here.

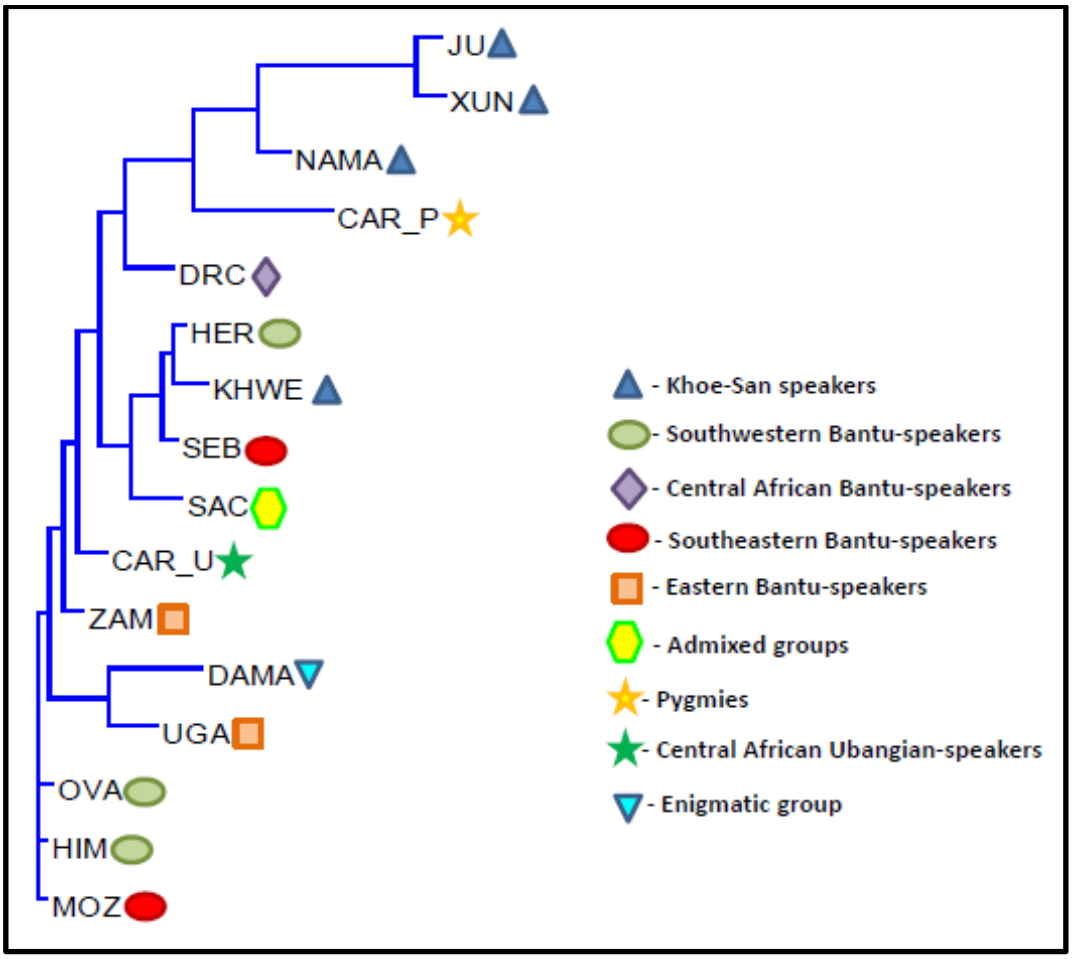

Figure 26 Neighbour-Joining population tree for indigenous populations

The spatial representation of the population's distances with each other is given in principle component analysis plots (PCA), Figure 27 and Figure 28. Figure 27 shows the PCA plot representing the majority of the variation detected, where the component 1 (x-axis) contains 64.9% of the total variation and the component 2 (y-axis) 18.3% of the variation. Relative to the distance observed on the x-axis: San populations JU and XUN cluster together and are furthest removed from other population; the Khoe (NAMA) is intermediately placed in

the plot along with the CAR_P; and all Bantu-speakers cluster together towards the left hand-side of the plot. There appears to be no geographic association among southeastern and Southwestern Bantu-speakers. Similarly, there is no geographic concordance between East, West, and Central African Bantu-peakers. Relative to the y-axis: Bantu-speakers appear somewhat different with DRC most removed from the group; again though no geographic concordance exists in this cluster. CAR_U cluster along with Bantu-speakers relative to the x-axis, and are further placed on the y-axis, possibly indicating an ancestral relation to all Niger-Kordafanian speakers where divergence had occurred over a long period. Relative to the x-axis the DAMA cluster closely with Bantu-speakers, but are further removed relative to the y-axis. DAMA though are much removed from neighbours NAMA, corroborating previous reports of this population more closely related to Bantu-speakers than their Nama dialect speaking neighbours (Nurse et al. ,1985; Soodyall and Jenkins, 1993; Soodyall et al., 2008; Schlebusch, 2010a). While NAMA and CAR_P are closes to one another relative to the x-axis, they are placed far apart on the y-axis.

Figure 28 shows the third dimension in the PCA plot, where the component 3 contains 12.7% of the total variation observed. Relative to component 3, all Bantu-speakers cluster tightly. Again no geographic concordance for Bantu-speakers appears to be evident. The CAR_U are positioned within this cluster, further affirming the previously mentioned ancestral relation to Bantu-speakers. Although the DAMA are closely positioned on the x-axis to Bantu-sepakers, relative to the third component are further removed. Khoe-San populations and CAR_P continue to be furthest removed from each other and Bantu-speakers in the plot. The SAC are too closely placed to Bantu-speakers, more closer than to Khoe-San populations. Their placement though needs to be treated with caution as their sea-borne parental populations SAW and SAI have been excluded here.

The results of the Analyses of Molecular Variance (AMOVA) for the indigenous African populations are summarized in Table 26. Populations were grouped by (1) Language, (2) Geographic position, and (3) Geography & Language. For all three AMOVA tests the DAMA and SAC were treated as separate entities. From the three AMOVA analyses the greatest variation can be found within populations. Test results for all permutations were all significant to $p<0.05$. Evident from the results of the three AMOVA tests performed, using language alone as a criterion to group populations (AMOVA analysis 1), the greatest proportion of variation was observed among groups rather than populations within the groups. This showing that language is a better confouder for these indigenous African populations.

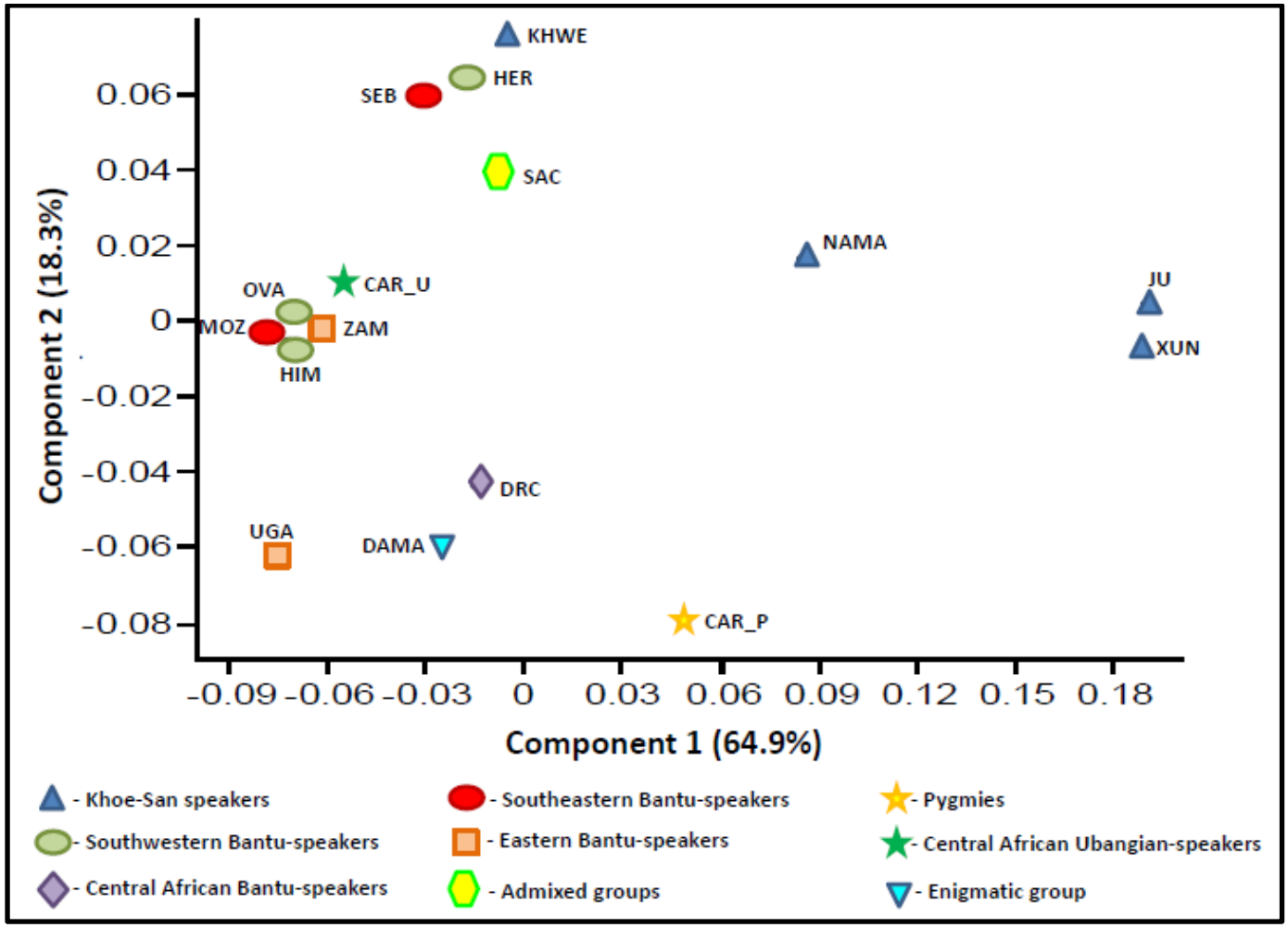

Figure 27 Principal component analysis of Fst values from allele distribution of 13 loci – component 1 versus component 2

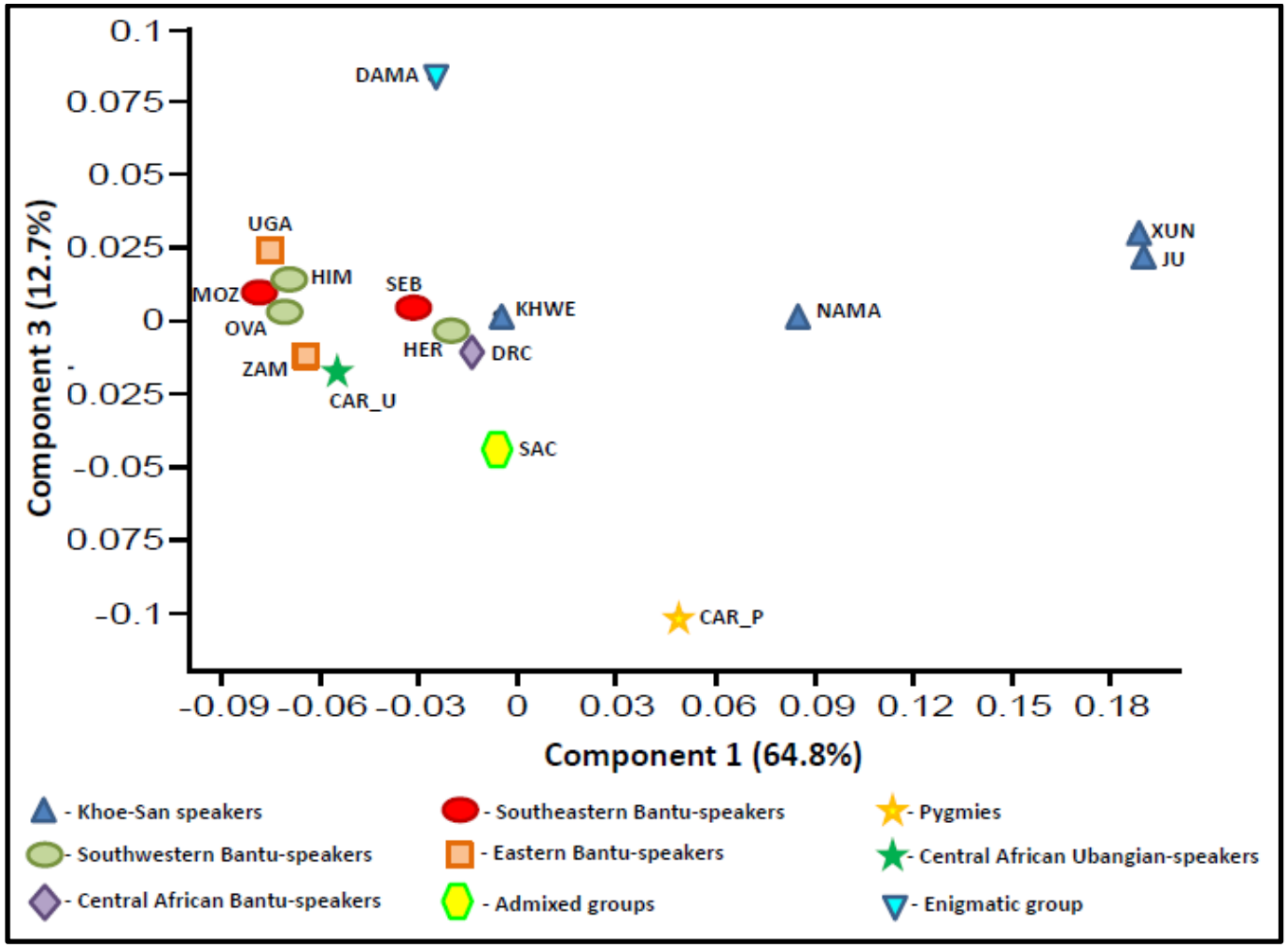

Figure 28 Principal component analysis of Fst values from allele distribution of 13 loci – component 1 versus component 3

Table 26 AMOVA analysis summary of indigenous African sample

AMOVA analysis number	Criteria for grouping	Grouping	Fst (% variation) Among groups	Among populations within a group	Within populations
1	Language	Khoe-San speakers (JU; XUN; KHWE; NAMA) Bantu-speakers (SEB; MOZ; HIM; HER; OVA; ZAM; UGA; DRC; CAR_U; CAR_P) Enigmatic (DAMA) English and Afrikaans speakers (SAC)	2.39	1.76	95.85
2	Geography	Southern Africa (JU; XUN; KHWE; NAMA; DAMA; SEB; HIM; HER; OVA; SAC) Eastern Africa (MOZ; ZAM; UGA) Central Africa (DRC; CAR_U; CAR_P)	1.10	1.89	97.01
3	Geography & Language	Southern African Khoe-San speakers (JU; XUN; KHWE; NAMA) Southeastern Bantu-speakers (SEB; MOZ) Southwestern Bantu-speakers (HIM; HER; OVA) Eastern Africa (ZAM; UGA) Central Africa (DRC; CAR_U; CAR_P) Enigmatic (DAMA) South African mixed population - English and Afrikaans speakers (SAC)	1.41	1.48	97.11

Note all values significant to $P<0.05$

3.7. Population structure

Allele frequencies for SNP loci investigated in this study were used in the software program Structure to assess the average genetic variation within the study population. Figure 29 depicts the results of the apportioned variation that exists individually and the average within a population (population assignment). Iterations for K=2 to K=6 were done, where K represents the number of portions or clusters the variation had been divided into. In the K=2 analysis, two portions of the genetic variation was assigned where the green portion appears larger in oldest African populations, the Khoe-San and the CAR_P. The green portion likely represents that which is ancient in humans, and the red that which is more recent in variation, which tends to be larger in Bantu-speakers, SAC, SAW and SAI. With successive runs K was increased, where the green and the red portions differentiate further. At K=6 the greatest likely average posterior value is observed (likely best number of clusters for data), where the green portion continues to be larger in Khoe-San and CAR_P populations compared to other populations. The green portion is at its lowest in SAC, and sea-borne migrants SAW and SAI. Further at K=6, the yellow portioned variation is smallest in most African populations appearing larger in SAW and SAI. With the exception of these differences, when the variation is apportioned into six clusters there appears to be little distinction between the apportion variation within each population.

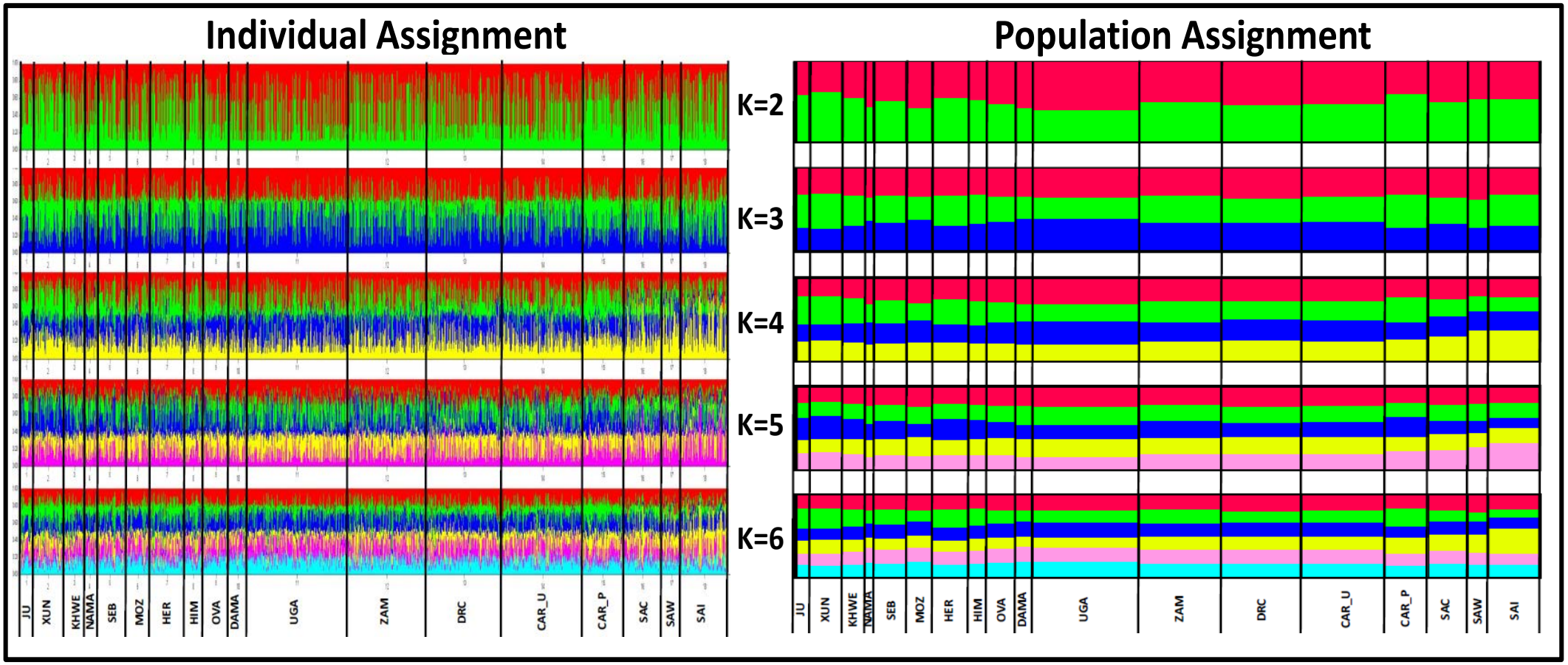

Figure 29 Results of the Structure runs (K=2 to K=6) for SNP loci in genes *CYP1A1, CYP1A2, CYP2A6, CYP2D6, CYP2E1, GSTP1* and *NAT2*

CHAPTER 4 DISCUSSION

The interaction of phase I and phase II metabolic genes are crucial to the maintenance of cell integrity and the normal process of cell proliferation, where if defaulted, could result in the formation of DNA adducts and cancer progression (Futreal et al., 2004). While many association studies investigating these genes have focused on European and Asian populations, little attention has been paid to the role of these genes among African populations. Given the paucity of genetic data from sub-Saharan Africa populations, this study was designed to unravel the genetic variation that exists in this region. The polymorphisms selected for the study were identified from the literature where loci had been reported to confer moderate susceptibility to cancers. While this study has no direct clinical relevance; the screen for 15 loci at nine genes provides a robust survey of genetic variation across the African continent for these important xenobiotic genes. The genetic relationships between 18 sub-Saharan African populations were assessed with respect to these loci. Data from literature was used to compare the genetic variation between Africans and non-Africans to identify differences in distribution of mutant alleles globally, and to assess their evolutionary histories. In deciphering the global distribution patterns of loci investigated, the study provides important insight into the relevance of these candidate loci for future studies on cancer and pharmacogenetics in Africa. To our knowledge this is a first most comprehensive study to investigate these cancer related genetic markers in Africa.

4.1. Phase I metabolizers

CYP1A1, CYP1A2, CYP2A6, CYP2D6 and *CYP2E1* genes have been for several years been implicated in cancer susceptibility (Lu et al., 1986; Watanabe et al., 1990;

Persson et al., 1993; Watanabe et al., 1994; Fernandez-Salguero et al., 1995; Mirvish et al., 1995; Hildesheim et al., 1997; Wu et al., 1998; Oscarson et al., 1998; Pianezza et al., 1998; Ingelman-Sundberg et al., 1999; Coutts and Urichuk, 1999; London et al., 1999; Nakajima et al., 1999; Olshan et al., 2000; Murata et al., 2001; Yoshida et al., 2002; Fujihara et al., 2009; Santovito et al.,2010; Piacentini et al., 2011). These genes activate carcinogens and other exogenous agents into more reactive chemical compounds which have been evident in cancer progression due to their DNA damage capability (Watanabe et al., 1994). Phase I mutations investigated in this study were CYP1A1*2A and *2C; CYP1A2*1C and *1F; CYP2A6*7 and *8; CYP2D6*3A (2549delA) and CYP2D6*4(1846G>A); and CYP2E1*5B(PstI) and CYP2E1*5B(RsaI).

CYP1A1*2A was prevalent in all global populations, to varying degrees in sub-Saharan Africans. Among study populations, the allele was most frequent in Khoe-San populations. Conversely, CYP1A1*2C allele was near absent in sub-Saharan Africans. Either mutant allele renders high enzyme activity of the *CYP1A1* gene (da Silva et al., 2008) which could result in the high levels of reactive oxidants from the metabolism of many carcinogens including polycyclic aromatic hydrocarbons. This could indicate that these populations are susceptible to lung, oesophageal, oral and liver cancers (Wu et al., 1998; Fujihara et al., 2009; Piacentini et al., 2011). The incidence of such cancers have been noted as highly prevalent in sub-Saharan Africans (Ferlay et al., 2010), hereby indicating possible genetic implications. The presence of both CYP1A1*2A and *2C mutant alleles is appreciably prevalent in SAW and SAI, consistent with previous reports of susceptibility to cancers (Watanabe et al., 1994).

Both CYP1A1*2A and CYP1A1*2C mutant alleles are more frequent in non-African populations, being most frequent in far-East Asians. Far-East Asians are likely to have many more high activity enzymes than Africans. This is consistent

with previous association study reports of susceptibility to many cancers in these populations (Wang et al., 2003; Kiruthiga et al., 2011; Xia et al., 2013; Yoshida et al., 2007; Islam et al., 2013; Souiden et al., 2012). High proportions for CYP1A1*2A and CYP1A1*2C mutant alleles are found in some Native American populations and could indicate possible susceptiblity to lung, oesophageal, oral and liver cancers.

CYP1A2*1C and *1F have both been associated in the aetiology of lung cancer, breast cancer, colorectal cancer and stomach cancer mainly in Europeans (Nakajima et al., 1999; Sachse et al., 1999; Hamdy et al., 2003b; Yoshida et al. 2007; B'chir et al., 2009; Wang et al., 2012). The study dataset contained large proportions of missing data for the CYP1A2*1C locus, and while the locus did not deviate from HWE for all populations, a more conservative consensus of its distribution in the study sample should be taken. CYP1A2*1C was found in all sub-Saharan African populations. Data from literature tends to suggest that CYP1A2*1C is most frequent in far-East Asians. CYP1A2*1F is found at high frequencies throughout the world, where in sub-Saharan African populations the mutation is present at frequencies greater than 40%. CYP1A2*1F‘s high prevalence in Khoe-San and CAR_P populations could indicate that the allele had evolved in Africa before humans migrated out of Africa.

The presence of CYP1A2*1F mutation decreases CYP1A2 enzyme activity (Cornelis et al., 2004). In contrast, the mutant allele for CYP1A2*1C renders an increased inducibility of the CYP1A2 enzyme (Cornelis et al., 2004). While little data was collected for the CYP1A2*1C locus, its presence of in most populations suggests that all populations in the study could contain *CYP1A2* genes that have increased inducibility properties.

CYP1A2 encodes for an enzyme that activates several compounds including carcinogenic arylamines, acetaminophen, caffeine and a number of widely prescribed antipsychotic drugs such as clozapine, imipramine, caffeine, paracetamol, phenacetin, theophylline, and tacrine (Wooding et al., 2002; Fujihara et al., 2007). It is highly expressed in the liver and lesser degree in the lung (Nakajima et al., 1999; B'chir et al., 2009). CYP1A2*1C high inducibility property induces highly active *CYP1A2* genes and could result in a greater conversion of the before mentioned compounds into reactive oxidant species, potentially creating greater stress on cell integrity and susceptibility to cancers. CYP1A2*1F decreased enzyme activity conversely potentially could result in a reduction in the conversion of these compounds and hereby possibly reduces the susceptibility to cancer formation. While CYP1A2*1F had not deviated from Tajima's D statistic test for selective neutrality, its decreased enzyme activity property and high prevalence could attribute to its retention in populations due to some protective measure. These functional profiles too would result in differences in *CYP1A2* metabolising drugs such as clozapine, imipramine, caffeine, paracetamol, phenacetin, theophylline, tracrine and some neurotoxins (Olshan et al., 2000).

Data for CYP2A6*7 and *8 mutations in the current sample and from the literature shows that these mutations are relatively underinvestigated. CYP2A6*7 was observed in most sub-Saharan African populations at varying frequencies, where the allele was most frequent in Central Africans. CYP2A6*7 was found in all Khoisan-speakers with the exception of NAMA and DAMA. Globally, CYP2A6*7 was most frequent in far-East Asians. In contrast, CYP2A6*8 was observed exclusively among Asian populations and possibly indicates an ascertainment bias where the mutation was first observed in far-East Asians.

The presence of the CYP2A6*7 mutant allele has been associated with reduced enzyme activity, while CYP2A6*8 is reported to render no effect (Landi et al., 1994; Parveen et al., 2010). The presence of the mutant allele for CYP2A6*7 in most populations in the study possibly indicates that most sub-Saharan populations contain low activity *CYP2A6* genes. *CYP2A6* genes have been implicated in reduced nicotine dependency and reduced risk to lung and oesophageal cancer (Nakajima et al., 1999; Santovito et al., 2010). There is though some debate to the association in nicotine dependency and reduced enzyme activity of CYP2A6 (Pianezza et al., 1998; London et al., 1999; Yoshida et al., 2002) and further investigation is necessary. This though could indicate the high presence of CYP2A6*7 in DRC could relate to a greater proportion of these individuals having reduced nicotine dependency. Likewise, DRC could be greatly limited in their ability in the uptake of pharmaceutical agents such as coumarin (+)-cis-3,5-dimethyl-2-(3-pyridl)thiazolidin-4-one hydrochloride, methoxy-flurane, halothane, losigamone, letrozole, valoproic acid, disulfiram, and fadrozole (Tan et al., 2001).

In this study, the CYP2D6*3A allele is absent in most Africans and minutely present in Europeans. CYP2D6*4 was found in almost all sub-Saharan Africans. With the exception of the KHWE the mutation was not found in Khoe-San populations. CYP2D6*4 is most frequent in Europeans, Indians and Middle Eastern populations; and absent in far-East Asians.

The presence of CYP2D6*4 in KHWE and CAR_P presents an interesting question: why is CYP2D6*4 absent in other Khoe-San populations? The population tree (Figure 26) suggests a very old ancestral relationship between Khoe-San and Pygmy (CAR_P). Firstly, it is possible that over time this mutantation was lost in Khoe-San due to genetic drift from these populations becoming extinct and living in isolation for many years. Secondly, it could be that it was introduced into the

KHWE and CAR_P by neighbouring Bantu-speaking populations. Another possibility could be that the mutant gene existed in Africa before the African exodus and following population divergence was retained in Bantu-speakers and possibly lost in Khoe-San. The last argument seems more plausible given the higher frequencies of CYP2D6*4 in Ghanaians, Tunisians and non-African populations; and the less likely that the mutation came from non-African sources. This argument could also explain the mutation's presence in the KHWE, where it is postulated that they were originally a Bantu-speaking group who changed their pastoral lifestyle for the hunter-gatherer lifestyle practised by neighbouring San. This further affirms previous reports of the evolutionary history of the KHWE. In assessing the genetic affinities of the populations in the study to each other (PCA plots and population tree), it is likely that the *CYP2D6* mutant gene was retained in the CAR_P, and lost in the Khoe-San.

The presence of either mutant alleles for CYP2D6*3A and CYP2D6*4 renders poor metabolism properties for the CYP2D6 enzyme (Rostami-Hodjegan et al., 1998; Arneth et al., 2009). The presence of these mutations in Europeans, and Middle Eastern populations is an indication that these populations possibly have many more poor-metabolizer *CYP2D6* genes in comparison with other populations. The high prevalence of CYP2D6*4 in Indians and Native American populations possibly indicates that these populations contain poor-metabolizer *CYP2D6* genes. With respect to the *CYP2D6* loci investigated here, the data suggests that poor-metabolizer genes are absent in far-East Asians with the exception of the Japanese who contain low proportion of the CYP2D6*4 mutation. However, it is possible other functional related mutations for the *CYP2D6* gene do exist in these populations and therefore cannot rule out the possible of dysfunctional *CYP2D6* genes in these populations. The presence of these poor-metabolizer genes possibly reduces the risk to lung cancer (Rostami-Hodjegan et al., 1998).

CYP2D6 gene is implicated in the metabolism of more than 25% clinically related drugs (van der Merwe et al., 2012), where reduced metabolic properties reduces the uptake of several neuroleptics, tricyclic antidepressant drugs, serotonin inhibitors and β-blockers (Linder et al., 1997; Sachse et al., 1997; Krajinovic et al., 1999). The reduction in uptake of such compounds is important in HIV drug treatment research (Brown et al., 2012). Breast cancer patients that have been prescribed tamoxifen (antidepressant) to cope with psychological effects of the condition have shown subsequent treatment relapses and associated increased risk of breast cancer recurring (van der Merwe et al., 2012), possibly due to dysfunctional *CYP2D6* genes. The data for the *CYP2D6* gene observed here suggests that most Khoisan-speakers could be susceptible to lung cancer and could likely metabolize drugs mentioned above better than other African populations.

Both CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I) loci are absent in sub-Saharan Africans and none contain mutant genes. *CYP2E1* genes containing both mutant alleles render a 10-fold increased enzyme activity (Watanabe et al., 1994; Kim et al., 1996; Bolt et al., 2003). Outside Africa, both mutations are present in far-East Asians and Chileans at high frequencies, and in Europeans and Indians at lower frequencies. By inference, it is possible that far-East Asians and Chileans have high frequencies of the 10-fold increased enzyme activity genes; being present at lower frequencies in in Europeans and Indians. It is therefore possible that far-East Asians, Europeans and Chileans could be genetically susceptible to nasopharyngeal, oesophageal, oral, prostate and lung cancer progression (Lu et al., 1986; Watanabe et al., 1990; Persson et al., 1993; Mirvish et al., 1995; Hildesheim et al., 1997; Sachse et al., 1997; Linder et al., 1997; Coutts and Urichuk, 1999; Ingelman-Sundberg et al., 1999; Murata et al., 2001; Ferreira et al., 2003; Yang et al., 2006). It is further possible that individuals in these populations carrying these mutations could be highly induced by ethanol and

isoniazid; and could possibly have a high uptake or metabolism of anaesthetics, analgesics, antipyretics, theophyllines, and chlorzoxazones (Flockhart, 2007).

4.2. Phase II metabolizers

The presence of null mutations of the *GSTM1 and GSTT1* genes reduces the function of the respective gene which causes reduced expression, diminishing inactivation of various carcinogen-intermediates thereby increasing oxidative stress and the risk of DNA adducts forming (Olshan et al., 2000). Many studies have found that with the inability of these mutations in detoxifying tobacco-related carcinogens and several environmental carcinogens, individuals presenting with these mutations are more likely 'at risk' of developing Head and Neck Squamous cell carcinoma (HNSCC), lung, bladder and other types of cancers (Olshan et al., 2000; Casson et al., 2003; Gajecka et al., 2005; Leichsenring et al., 2006; Losi-Guembarovski et al., 2008; Torresan et al., 2008; Varela-Lema et al., 2008).

Within Africa the GSTM1*0 genotype was found in all populations albeit at lower frequencies compared to non-Africans. The mutation was though not present in JU. In Africa, the mutation increases in frequency from south to north where it is most frequent outside Africa. It is likely that GSTM1*0 was present in Africa long before human migration out of Africa, and over time due to founder effects during the course of migration throughout the world became highly prevalent in Europe, Asia and Middle East. Khoisan-speakers display the lowest frequency of GSTM1*0 among African populations, exceptions found in the KHWE and DAMA populations. Similar to trends seen for other loci, the GSTM1*0 mutation presence in the KHWE and DAMA is an indication of their genetic similarity to Bantu-speaking populations. This confirming previous reports of the genetic

histories of these population (Nurse et al., 1985; Soodyall et al., 2008; Henn et al., 2008; Schlebusch, 2010a; Schlebusch et al., 2012).

In contrast, GSTT1*0 genotype is highly prevalent in Africans and where there appears to be little or no clinal distribution in Africa as that seen for GSTM1*0. While, GSTT1*0 is found at appreciably high frequencies in the Khoe-San, it is very low in the CAR_P. Given the evolutionary history observed for these populations in this study, the low proportion of the GSTT1*0 genotype could be due to drift and bottleneck effects following decline in population size. Conversely, the GSTM1*0 genotype is found at higher frequencies in CAR_P than in Khoe-San population. Given the well documented evolutionary history of humans, the presence of both mutations in these populations could indicate their existence in Africa for thousand's of years, possible long before the African exodus. The proportions of both GSTM1*0 and GSTT1*0 null genotypes in populations of admixed ancestry (SAC, African-Americans and Brazilians) tend to be similar. Even though their ancestral histories are slightly different they have similar contributors in European and African descendants.

Almost all global populations display >20% proportions of either GSTM1*0 and GSTT1*0 null genotypes. Given that these mutations render an inability to detoxify tobacco-related carcinogens and xenobiotics, almost all global populations could be adversely predisposed to cancers. This predisposition is due to their inability in detoxifying reactive oxidant species produced by phase I metabolic enzymes or detoxifiers. Pharmacologically, GSTM1*0 and GSTT1*0 mutations have found to have significant association in hepatotoxicity induced by TB drugs (Roy et al., 2001; Huang et al., 2007; Leiro et al., 2008). However, one study has found no association in Indians (Chatterjee et al., 2010). Therefore further investigation to understanding the drug detoxifying effects of GSTM1*0 and GSTT1*0 mutations is required. The presence of these mutations found in

sub-Saharan African populations highlights the importance in understanding genetic compositions of population with relation to drug treatment. This would be vital in regard to the treatment of TB that is a major communicable disease concern in southern Africa.

There were individuals who were homozygous for both GSTM1*0 and GSTT1*0 mutations. Such individuals were found in all study populations, except the JU and OVA. The frequencies of individuals homozygous for both mutations were highest in Bantu-speakers and low in the Khoe-San and CAR_P. Where reported in the literature, non-Africans have higher frequencies of both mutations being present (Adams et al., 2003; Gundacker et al., 2007; Al-Dayel et al., 2008; Fujihara et al., 2009; Santovito et al., 2010; Salem et al., 2011). Individuals homozygous for both mutations effectively have both *GSTM1* and *GSTT1* non-functional genes. Given the effects of each mutation, these individuals are possibly more susceptible to cancer development, and likely lack the ability to metabolise exogenous substances.

Glutathione S-transferase class π (*GSTP1*) is an important phase II metabolising enzyme involved in detoxification of many hydrophobic and electrophilic compounds by catalytic conjugation activity. In this study two mutations for this gene was studied, GSTP1*Ile105Val and GSTP1*Ala114Val. The GSTP1*Ile105Val mutant allele was found in all African and other global populations. This mutation is frequently evident in southern African populations and found in high frequencies in Khoisan-speakers. GSTP1*Ile105Val decreases in frequency from the south towards north Africa, with lower frequencies observed in European, Asian and American populations. The GSTP1*Ala114Val mutant allele in contrast is near absent in sub-Saharan Africans.

The presence of the mutant allele for GSTP1*Ile105Val is reported to render a 3-fold decreased enzyme activity (Ali-Osman et al., 1997). This mutation is frequent in most populations of this study, and most prevalent in the Khoe-San. The presence of either GSTP1*Ile105Val and GSTP1*Ala114Valalleles has been reported to render a decrease in enzyme activity (Ali-Osman et al., 1997). The high frequencies of both these alleles in SAW and SAI is consistent with observations made in European and Asian populations, where Europeans and Asians contain a large proportion of *GSTP1* genes with 3 to 4-fold reduced activity (van Lieshout et al., 1999; Abbas et al., 2004; Buchard et al., 2007).

This reduced enzyme activity properties have for many years been associated with susceptibility to many head and neck cancers, including oesophageal and lung cancer (Cho et al., 2005); and associated reduced ability to metabolize chemotherapeutic agents, including melphalan, cyclophosphamide, vincristine, adriamycin, cisplatin, etoposide, thiotepa, chlorambucil, and busulphan (Czerwinski et al., 1996; Awasthi et al., 1996). The presence of GSTP1*Ile105Val in sub-Saharan Africans (more so Khoisan-speakers), possibly indicates that a few individuals containing this mutation from these populations are incapable of using their *GSTP1* genes in detoxifying carcinogens and have predisposition to cancer development. Further, the reduced ability to metabolize chemotherapeutic compounds could result in an altered susceptibility to chemotherapy-induced carcinogenesis and indicates the need for careful drug selection in cancer therapy of these populations. Allan et al. (2001) showed that the GSTP1*Ile105Val mutant allele was associated with reduced activity and inability to detoxify chemotherapy drugs, specifically to therapy-related acute myeloid leukemia, which can produce devastating complications in long-term cancer survival.

NAT2*14A, consistent with previous reports (Dandara et al., 2003; Teixeira et al., 2007) was observed exclusively in Africans and admixed populations of African descent. The *NAT2* gene is responsible for the N-acetylation of certain aryl amines, xenbiotics and tobacco-related carcinogens. The polymorphism NAT2*14A renders poor metabolising function and is likely to render associated risk in African populations in the development of urinary bladder, colorectal, breast, head and neck, lung and prostate cancers (Hein, 2000; Butcher et al., 2002). In anti-TB treatment, *NAT2* is associated with N-acetylation of Isoniazids (INH), one of the most important drugs used in TB chemoprophylaxis (Teixeira et al., 2007). TB is highly prevalent in 3.9 million Africans and with an incidence rate in Africa of 340 person per 100,000 a year is responsible for 50/100,000 deaths per year (http://www.who.int/mediacentre/factsheets/fs104/en/). The poor metabolic properties of NAT2*14A in Africans thus becomes of particular concern in anti-TB treatment choice. This result suggests a likely reason for INH treatment failure in some sub-Saharan Africans. However, there are four other mutations on this gene which also have slow modulation properties which can be found in African (Dandara et al., 2003), therefore illustrates the need to investigate *NAT2* polymorphisms further in African populations.

4.3. Role of Phase I and Phase II metabolisers in Africa

In humans the interaction of phase I metabolisers with phase II metabolisers play an important role in maintaining cell integrity and preventing cancer progression (Furberg and Ambrosone, 2001). Bartsch et al. (2000) highlighted the importance of *CYP1A1* mutations in combination with GSTM1 null mutations in smoking related cancers. Similarly, the interaction of *CYP2E1* and *GSTP1* genes have been implicated in increased risk of cancer development to European industrial workers exposed to styrene (Thier et al., 2003). Another example where toxic nitrosamines found in tobacco smoke can be slowly degraded overtime in

individuals who contain slow modulators for the *NAT2* gene, in combination with high acetylators in the *CYP2D6* and *CYP1A2* genes could result in increased risk to cancer (Nebert, 1997). Mutations of *CYP1A1*, *CYP2A6* and *CYP2E1*, along with the deficiencies of GST genes have long been associated with increased susceptibility to various cancers (Dandara et al., 2001).

These studies are a few examples of the interaction of phase I and II genes necessary to maintain cell integrity. These associations have come from several studies done in European and Asian populations, and associated with different environmental factors. Given that cancer is a multifactorial disease we cannot exclude the possibility that there are other biological and environmental variables which contribute to these associations, which are possibly not present in sub-Saharan Africans. However, if these interactions are present in sub-Saharan Africans what can be deduced from the data of the present study?

In review of the genetic data in this study alone, given the association identified in Bartsch et al. (2000), it is possible that this association may not be relevant to Africans. Africans have a higher predominance for the GSTT1*0 null mutation than for the GSTM1*0, and the CYP1A1*2C mutation is near absent. Thier et al. (2003) implicated CYP2E1 and GSTP1 mutations in cancer progression. Given the absence of CYP2E1 mutations in the study, this combined genetic factor cannot be confirmed for indigenous Africans. The association between slow modulators for the *NAT2* gene, in combination with high acetylators in the *CYP2D6* and *CYP1A2* genes (Nerbet, 1997), could possibly be demonstrated in study populations. Africans contain the NAT2*14A allele rendering poor metabolic properties for the *NAT2* gene and have the CYP1A2*1C mutation which renders an increased inducibility of the CYP1A2 enzyme. These genetic combinations though cannot be entirely confirmed or refuted though in this study. Therefore,

such combinations should be investigated further in African populations along with associated risk to other biological and environmental factors.

4.4. Population genetic diversity

Tajima's D statistic was applied to all 13 SNP loci in each population to assess if they were under neutral selection. While several loci were found to deviate from HWE within certain populations, only CYP1A1*2C was found to be under selection within the XUN population albeit that the locus was not found to deviate from HWE. In addition, an analogous computation to the HWE test, the exact test was applied to determine population differences. Unlike the HWE, the exact test indicates a larger proportion of populations that were significantly different when compared to each other in pair-wise comparisons.

A possible reason for the differences in the outcome of the HWE and exact tests could be due to the varying sample sizes. The exact test is known to be more accurate when the expected number is low (Excoffier et al., 2005). In the case of the study sample, some populations had smaller sample sizes to others. It is likely that populations contained lower proportions of alleles than the expected outcomes. The latter is observed in the low average gene diversity and low average observed heterozygosity seen in each population across the 13 SNP loci investigated. These low diversity indices observed is likely due to all populations being monomorphic for CYP2A6*8, CYP2D6*3A, CYP2E1*5B(*Pst*I) and CYP2E1*5B(*Rsa*I) loci, accounting for 30.8% of the loci investigated in the study. In addition to these loci, indigenous African populations in the sample had many other monomorphic loci. CYP1A1*2C loci was monomorphic in all populations with the exception of the XUN, MOZ and HIM populations; likely due to recent introduction of the mutant allele from non-African sources. Similarly

GSTP1*Ala114Val loci was monomorphic in almost all populations, except for DRC, HER, MOZ, SEB and UGA populations where the proportions of the mutant allele is low and likely came from non-African individuals. CYP2A6*7 loci was monomorphic in the NAMA, DAMA and OVA populations. The CYP2D6*4 loci was monomorphic in near all Khoe-San populations (JU, XUN and NAMA) and HIM population. In contrast, sea-borne migrants SAW and SAI had 7 polymorphic loci, and the admixed SAC had 8 polymorphic loci.

The presence of these monomorphic loci in the indigenous African populations is not only key in understanding the evolutionary histories of these populations, but enhances our understanding of the genetic predisposition to cancers in these populations. All monomorphic loci highlighted above contained the wild-type allele. Therefore it is possible that these monomorphic loci are not candidate loci to be used for future cancer clinical studies within African populations. Given the increasing concern of the cancer burden in Africa, other markers implicated in cancer susceptibility should be investigated to better understand the genetic susceptibility that exists among Africans.

4.5. Genetic relationships of sub-Saharan Africans

To elucidate the genetic relationship of sub-Saharan Africans in the study, the genetic distances of indigenous African populations were used in PCA, population tree and AMOVA analyses. The results of these analyses show that sub-Saharan Africans are largely similar to each other, with a few exceptions. The PCA plot indicates that Bantu-speakers are closely related with little or no difference that corresponds to their geographic positions. The placement of Bantu-speakers in the topology of the population tree confirms this. AMOVA tests further indicate the greatest population difference is between language

groups, and not for populations grouped to their geographical locations. The great Bantu expansion is more likely the reason for this observation where migrants practicing pastoral and agricultural lifestyles left West Africa to populate the rest of sub-Saharan Africa during a period spanning approximately 4000 years (Nurse et al., 1985; Güldemann et al., 2002; Lambert and Tishkoff, 2009). Bantu-speakers moved to southern Africa following both eastern and western routes ~1500 years ago (Nurse et al., 1985). It is possible that time depth of these dispersion events would have been too short to allow for greater genetic differentiation. The Khoe-San were the aboriginal populations of southern Africa and they too remained genetically distinct from their Bantu-speaking neighbours. Hence, language was evident as a major confounder to the variation seen among sub-Saharan Africans in this study. Previous reports of African populations concordant to language relationships corroborates these findings (Nurse et al., 1985; Cavalli-Sfroza et al., 1994; Tishkoff et al,. 2007; Behar et al., 2008; Naidoo et al., 2010; Schlebusch et al., 2010b; Schlebusch et al., 2012). Language is reported to have a near one-to-one relationship with "tribal" or ethnicity grouping and where linguistic classification is synonymously used to define a population groups (Cavalli-Sforza et al., 1994; MacEachern et al., 2000). Similarly, a putative parallelism of linguistics and genetic change exists in Africans (MacEachern et al., 2000; Scheinfeldt et al., 2010).

When migrating to the south, Bantu-speakers would have encountered Khoisan-speakers. During these encounters Bantu-speakers would have introduced culture and subsistence practises to the Khoe-San populations (Nurse et al., 1985). The interactions between these populations resulted largely in asymmetric assimilation of Khoe-San women into Bantu-speaking populations (Tishkoff et al., 2007; Soodyall et al., 2008; Naidoo et al., 2010; Schlebusch et al., 2011; Sikora et al., 2011). This aspect could explain the similarities between some Khoisan-speaking populations and Bantu-speaking populations. However,

in the case of the KHWE, their genetic similarities to Bantu-speakers are much greater than to Khoisan-speakers. This finding is consistent with previous reports that the KHWE are not true Khoe-San. That they may be descendants of an east African pastoralist group that introduced the pastoralist culture into present day northern Botswana (Henn et al., 2008; Schlebusch et al., 2010a; Schlebusch et al., 2012). In contrast, the DAMA appear genetically closer to Bantu-speakers while speaking the Nama dialect. This relationship further confirms the proposed genetic heritage of DAMA being of Bantu-speaking groups (Nurse et al. 1985; Soodyall et al., 2008; Schlebusch et al., 2010a; Schlebusch et al., 2012).

Among Khoisan-speakers, the JU and XUN cluster together and are far removed from other clusters in the PCA plot. The population tree confirms this relationship and indicates these populations being the oldest in the sample. The NAMA too are quite different from other populations where their placement in the PCA is somewhat intermediate between Khoe-San and Bantu-speakers. The placement of the NAMA and the branch length in the population tree indicates NAMA are closely related to JU and XUN populations, but not as old. The CAR_P too show to be one of the older populations and distantly related to Bantu-speakers and the Khoe-San. These differences are consistent with previous reports of the JU, XUN and CAR_P populations representing some of the oldest surviving populations in Africa (Nurse et al. 1985; Sikora et al., 2011; Schlebusch et al., 2012). All three populations followed hunter-gatherer subsistence and for many years due to linguistic and anthropological barriers had little interaction with neighbouring Bantu-speakers (Tishkoff et al., 2007; Nurse et al., 1985; Sikora et al., 2011; Schlebusch et al., 2012).

The South African Coloured represent an admixed group born primarily from the unions of individuals from Khoisan-speaking, Bantu-speaking, South African White populations, and to a degree from Asian populations. Data in the study

indicates a high degree of mutant alleles not normally present in Africans having come from sea-borne migrants. The SAC population appeared to have closer genetic affinities to Bantu-speakers than to the Khoe-San. The SAC were sampled in the Gauteng province, South Africa. They had somewhat different histories to other coloured groups of South Africa where considerable contact with Bantu-speakers were likely. This in comparison to Coloured groups of the Western Cape who have had considerable contributions from the Khoe-San or are extinct populations of the KhoeKhoe misclassified as 'Coloured' (Schlebusch, 2010a). The different histories of these South African Coloured groups could explain the differences in mutation proportions observed in the study between the SAC and Coloured groups from literature sources.

Among the Eurasians, the SAW and SAI are largely distinct from sub-Saharan Africans. These populations mostly contained mutations that were present in parental populations and absent in their southern African neighbours. This aspect is confirmed by the higher than average gene diversity observed for these populations. For certain loci SAW and SAI contained slightly lower proportions of mutant alleles in comparison to parental populations. This trend is likely due to founder effects following their migration to South Africa, where their gene pools were a subset of that found in parental populations and the current surviving individuals have maintained these allele proportions.

Structure analysis further confirmed these genetic relationships observed in PCA and the population tree. In apportioning the variation observed in the sample into 'K' clusters, the highest likelihood of the divisions was reached at K=6. At K=6 the JU, XUN and CAR_P contained higher proportions of older clusters. Bantu-speakers contained similar proportions of the six clusters. The KHWE and DAMA too had similar proportions of the sixth cluster as that of Bantu-speakers.

Sea-borne migrants contained a larger proportion of the sixth cluster, further distinguishing these populations from Africans in the study.

4.6. Limitations of the study

Caution should be taken when reviewing the evolutionary associations observed here and the implication of these variants in disease susceptibility to cancer. Similarly, no direct conclusions can be drawn to these variant's pharmacogenetic properties in the study sample. The resolution within these genes is only a subset of those known to exist. More extensive research is required. Since no pathology information was collected from volunteers at the time of sample collection understanding these gene's association to disease cannot be accurately inferred. Further, there could be other biological and environmental factors that could not be considered in this study which may impact on the underlying relevance of these gene mutations to diseases and pharmacogenetic properties.

The low diversity observed in many of these populations could likely explain the largely compactness of the relationships detected. Therefore these findings should be seen as that representative of the markers investigated here, and that they do not necessarily relate to actual genetic distances between these populations. Where corroborating evidence from literature has been highlighted here, it more likely affirms the trends in evolutionary history seen among these populations.

4.7. Future studies and concluding remarks

While the study has improved our understanding of the distribution of these mutations in sub-Saharan Africans, further investigation is required. Only a subset of the loci implicated in cancers was investigated and the monomorphic loci found in indigenous Africans is an indicator of these loci being less informative in an African context to assess cancer susceptibility. Further, there are several other biological factors to consider and therefore the interpretation to the relative genetic predisposition from this study should be treated with caution.

In designing future studies, the evolutionary biology of these phase I and phase II metabolism genes and their functional interactions needs to be taken into account. Such genes have evolved over millions of years, in eukaryotes and other multicellular organisms to render vital life functions. Overtime these genes would have mutated towards adaptation in diets, niches and exposure to various xenobiotic species. The divergence from eukaryotes to multicellular organisms would have required dynamic interactions of these gene products (Nebert, 1997). It is possible various polymorphism had been retained in populations as their combinations allowed for balanced functionality. This “balanced trait” would have allowed for ‘shared benefit’ in various metabolic systems and thereby maintaining functionality to survive during African’s exposure to changing climatic and dietary conditions as they migrated throughout Africa.

Another aspect to consider in the increasing incidence of cancer in Africans, is that Africans are fast adopting western lifestyles (Ferlay et al., 2010; Somdyala et al., 2010; Wentink et al., 2010; Parkin et al., 2010; Dalal et al., 2011; Jemal et al., 2012; Chokunonga et al., 2013). Genetic adaptations to environmental pressures is less likely in the next few decades and necessitates that future studies

investigate these environmental factors along with genetic factors in identifying associated risk factors to cancer that presents in Africa. Therefore, future studies should take cognisance to incorporating larger sample sizes, increase the number of genetic loci to investigate, and where possible incorporate other biological and environmental factors. By this design, future studies may unravel the mysteries in cancer progression within Africa with greater rigour.

With respect to these loci we confirm previous genetic relationships reported for study populations. The study has shown the importance in elucidating the ancestral backgrounds of African populations to understand the possible underlying molecular epidemiology of certain diseases, namely cancer. More specifically, using a population genetic approach the study has shown that linguistic and anthropological relationships should be considered in the design of future clinical trials within sub-Saharan Africa.

The data further indicated candidate loci which could be used in future clinical trial studies to better understand cancer progression in Africa. Non-communicable diseases like cancer are on the increase in Africa, exacerbated by the increasing adoption to western lifestyles by Africans. Future studies would probably identify more candidate loci, and by incorporating other association factors the underlying aetiology of cancers and other non-communicable diseases may become evident. Finally, by unravelling the global distribution of mutations using both current study data and that found in the literature, the study has added to the pharmacological picture with respect to loci investigated, which could have benefits to the design of future pharmacological studies.

REFERENCES

Abbas A, Delvinquiere K, Lechevrel M, Lebailly P, Gauduchon P, Launoy G, et al. GSTM1, GSTT1, GSTP1 and CYP1A1 genetic polymorphisms and susceptibility to esophageal cancer in a French population: different pattern of squamous cell carcinoma and adenocarcinoma. World J Gastroenterol. 2004;10:3389–3393.

Abdel-Rahman HA, Khorshied MM, Elazzamy HH, Khorshid OM. The link between genetic polymorphism of glutathione-S-transferases, GSTM1, and GSTT1 and diffuse large B-cell lymphoma in Egypt. J Cancer Res Clin Oncol. 2012 Aug;138(8):1363-8.

Adams CH, Werely CJ, Victor TC, Hoal EG, Rossouw G, van Helden PD. Allele frequencies for glutathione S-transferase and N-acetyltransferase 2 differ in African population groups and may be associated with oesophageal cancer or tuberculosis incidence. Clin Chem Lab Med. 2003;41(4):600-5.

Agorastos T, Papadopoulos N, Lambropoulos AF, Chrisafi S, Mikos T, Goulis DG, Constantinidis TC, Kotsis A, Bontis JN. Glutathione-S-transferase M1 and T1 and cytochrome P1A1 genetic polymorphisms and susceptibility to cervical intraepithelial neoplasia in Greek women. Eur J Cancer Prev. 2007;16(6):498-504.

Agorastos T, Papadopoulos N, Lambropoulos AF, Chrisafi S, Mikos T, Goulis DG, Constantinidis TC, Kotsis A, Bontis JN. Glutathione-S-transferase M1 and T1 and cytochrome P1A1 genetic polymorphisms and susceptibility to cervical intraepithelial neoplasia in Greek women. Eur J Cancer Prev. 2007;16(6):498-504.

Aklillu E, Carrillo JA, Makonnen E, Hellman K, Pitarque M, Bertilsson L, et al. Genetic polymorphism of CYP1A2 in Ethiopians affecting induction and

expression: characterization of novel haplotypes with single-nucleotide polymorphisms in intron 1. Mol Pharmacol. 2003;64(3):659-69.

Al-Dayel F, Al-Rasheed M, Ibrahim M, Bu R, Bavi P, Abubaker J, et al. Polymorphisms of drug-metabolizing enzymes CYP1A1, GSTT and GSTP contribute to the development of diffuse large B-cell lymphoma risk in the Saudi Arabian population. Leuk Lymphoma. 2008;49:122–129.

Ali-Osman F, Akande O, Antoun G, Mao JX, Buolamwini J. Molecular cloning, characterization, and expression in Escherichia coli of full-length cDNAs of three human glutathione S-transferase Pi gene variants. Evidence for differential catalytic activity of the encoded proteins. J Biol Chem. 1997;272(15):10004-12.

Allan JM, Wild CP, Rollinson S, Willett EV, Moorman AV, Dovey GJ, et al. Polymorphism in glutathione S-transferase P1 is associated with susceptibility to chemotherapy-induced leukemia. Proc Natl Acad Sci. 2001;98(20):11592-7.

Al-Moundhri MS, Al-Kindi M, Al-Nabhani M, Al-Bahrani B, Burney IA, Al-Madhani A, et al. NAT2 polymorphism in Omani gastric cancer patients-risk predisposition and clinicopathological associations. World J Gastroenterol. 2007;13(19):2697-702.

Altayli E, Gunes S, Yilmaz AF, Goktas S, Bek Y. CYP1A2, CYP2D6, GSTM1, GSTP1, and GSTT1 gene polymorphisms in patients with bladder cancer in a Turkish population. Int Urol Nephrol. 2009;41(2):259-66.

Amer MA, Ghattas MH, Abo-Elmatty DM, Abou-El-Ela SH. Influence of glutathione S-transferase polymorphisms on type-2 diabetes mellitus risk. Genet Mol Res. 2011 Oct 31;10(4):3722-30.

Anderson LN, Cotterchio M, Mirea L, Ozcelik H, Kreiger N. Passive cigarette smoke exposure during various periods of life, genetic variants, and breast cancer risk among never smokers. Am J Epidemiol. 2012;175(4):289-301.

Arneth B, Shams M, Hiemke C, Härtter S. Rapid and reliable genotyping procedure for detection of alleles with mutations, deletion, or/and duplication of the CYP2D6 gene. Clin Biochem. 2009;42(12):1282-90.

Arvanitidis K, Ragia G, Iordanidou M, Kyriaki S, Xanthi A, Tavridou A, Manolopoulos VG. Genetic polymorphisms of drug-metabolizing enzymes CYP2D6, CYP2C9, CYP2C19 and CYP3A5 in the Greek population. Fundam Clin Pharmacol. 2007;21(4):419-26.

Awasthi S, Srivatava SK, Piper JT, Singhal SS, Chaubey M, Awasthi YC. Curcumin protects against 4-hydroxy-2-trans-nonenal-induced cataract formation in rat lenses. Am J Clin Nutr. 1996 Nov;64(5):761-6.

Aydin-Sayitoglu M, Hatirnaz O, Erensoy N, Ozbek U. Role of CYP2D6, CYP1A1, CYP2E1, GSTT1, and GSTM1 genes in the susceptibility to acute leukemias. Am J Hematol. 2006 Mar;81(3):162-70.

Aynacioglu AS, Cascorbi I, Mrozikiewicz PM, Roots I. Arylamine N-acetyltransferase (NAT2) genotypes in a Turkish population. Pharmacogenetics. 1997;7:327-331.

Badro DA, Douaihy B, Haber M, Youhanna SC, Salloum A, Ghassibe-Sabbagh M, Johnsrud B, Khazen G, Matisoo-Smith E, Soria-Hernanz DF, Wells RS, Tyler-Smith C, Platt DE, Zalloua PA; Genographic Consortium. Y-chromosome and mtDNA genetics reveal significant contrasts in affinities of modern Middle Eastern populations with European and African populations. PLoS One. 2013;8(1):e54616.

Balaji L, Singh KB, Bhaskar LV. Genetic polymorphisms of the CYP2E1 gene do not contribute to oral cancer susceptibility in south Indians. Asian Pac J Cancer Prev. 2011;12(6):1523-7.

Ballerini S, Bellincampi L, Bernardini S, Iori R, Cortese C, Federici G. Analysis of GSTP1-1 polymorphism using real-time polymerase chain reaction. Clin Chim Acta. 2003;329(1-2):127-32.

Barbieri C, Butthof A, Bostoen K, Pakendorf B. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur J Hum Genet. 2013;21(4):430-6.

Barkhan D. Haploid genetic variation in populations from Uganda, Zambia and the Central African Republic. University of the Witwatersrand. 2004.

Barnard A (1992). Hunters and herders of southern Africa - A comparitive ethnography of the Khoisan peoples. Cambridge, Cambridge University Press.

Bartsch H, Nair U, Risch A, Rojas M, Wikman H, Alexandrov K. Genetic polymorphism of CYP genes, alone or in combination, as a risk modifier of tobacco-related cancers. Cancer Epidemiol Biomarkers Prev. 2000;9(1):3-28.

Bayoumi RA, Qureshi MM, al-Ameri MM, Woolhouse NM. The N-acetyltransferase G191 A mutation among Sudanese and Somalis. Pharmacogenetics. 1997 Oct;7(5):397-9

B'chir F, Pavanello S, Knani J, Boughattas S, Arnaud MJ, Saguem S. CYP1A2 genetic polymorphisms and adenocarcinoma lung cancer risk in the Tunisian population. Life Sci. 2009;84(21-22):779-84.

Beck E. The Historisch-Geographische Pathologie of August Hirsch. A contribution from the 19th century to the pathomorphology of diseases. Gesnerus. 1961;18:33-44.

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Makkan H, et al. The dawn of human matrilineal diversity. Am J Hum Genet. 2008;82(5):1130-40.

Bhana S, Brain JB (1990) Setting down roots: Indian Migrants in South Africa, 1860-1911. Witwatersrand University Press, Johannesburg

Bolt HM, Roos PH, Thier R. The cytochrome P-450 isoenzyme CYP2E1 in the biological processing of industrial chemicals: consequences for occupational

and environmental medicine. Int Arch Occup Environ Health. 2003;76(3):174-85.

Bradford LD. CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. Pharmacogenomics. 2002;3(2):229-43.

Brauer G. The Origin of Modern Anatomy: By Speciation or Intraspecific Evolution. Evolutionary. Anthropology 2008;17:22–37.

Brower V. AIDS-related cancers increase in Africa. J Natl Cancer Inst. 2011;103(12):918-9.

Brown KC, Hosseinipour MC, Hoskins JM, Thirumaran RK, Tien HC, Weigel R, Tauzie J, Shumba I, Lamba JK, Schuetz EG, McLeod HL, Kashuba AD, Corbett AH. Exploration of CYP450 and drug transporter genotypes and correlations with nevirapine exposure in Malawians. Pharmacogenomics. 2012;13(1):113-21.

Bu R, Gutiérrez MI, Al-Rasheed M, Belgaumi A, Bhatia K. Variable drug metabolism genes in Arab population. Pharmacogenomics J. 2004;4:260-266.

Buchard A, Sanchez JJ, Dalhoff K, Morling N. Multiplex PCR detection of GSTM1, GSTT1, and GSTP1 gene variants: simultaneously detecting GSTM1 and GSTT1 gene copy number and the allelic status of the GSTP1 Ile105Val genetic variant. J Mol Diagn. 2007;9:612–617.

Butcher NJ, Boukouvala S, Sim E, Minchin RF. Pharmacogenetics of the arylamine N-acetyltransferases. Pharmacogenomics J. 2002;2(1):30-42.

Buzková H, Pechandová K, Slanar O, Perlík F. Frequency of single nucleotide polymorphisms of CYP2D6 in the Czech population. Cell Biochem Funct. 2008;26(1):76-81.

Campbell MC, Tishkoff SA. The evolution of human genetic and phenotypic variation in Africa. Curr Biol. 2010;20(4):R166-73.

Canova C, Richiardi L, Merletti F, Pentenero M, Gervasio C, Tanturri G, Garzino-Demo P, Pecorari G, Talamini R, Barzan L, Sulfaro S, Franchini G, Muzzolini C,

Bordin S, Pugliese GN, Macrì E, Simonato L. Alcohol, tobacco and genetic susceptibility in relation to cancers of the upper aerodigestive tract in northern Italy. Tumori. 2010;96(1):1-10.

Cantor KP, Villanueva CM, Silverman DT, Figueroa JD, Real FX, Garcia-Closas M, et al. Polymorphisms in GSTT1, GSTZ1, and CYP2E1, disinfection by-products, and risk of bladder cancer in Spain. Environ Health Perspect. 2010 Nov;118(11):1545-50.

Cascorbi I, Brockmöller J, Mrozikiewicz PM, Müller A, Roots I. Arylamine N-acetyltransferase activity in man. Drug Metab Rev. 1999;31(2):489-502.

Casson AG, Zheng Z, Chiasson D, MacDonald K, Riddell DC, Guernsey JR, et al. Associations between genetic polymorphisms of Phase I and II metabolizing enzymes, p53 and susceptibility to esophageal adenocarcinoma. Cancer Detect Prev. 2003, 27(2):139-46.

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

Cavalli-Sforza LL. African Pygmies: an evaluation of the state of research. Cavalli-Sforza LL (ed) African Pygmies. Academic Press inc., New York, 1986; Chapter 26:361-426.

Cavaco I, Reis R, Gil JP, Ribeiro V. CYP3A4*1B and NAT2*14 alleles in a native African population. Clin Chem Lab Med. 2003 Apr;41(4):606-9.

Cavaco I, Asimus S, Peyrard-Janvid M, Ferreira PE, Veiga MI, Hai TN, Ribeiro V, Ashton M, Gil JP. The Vietnamese Khin population harbors particular N-acetyltransferase 2 allele frequencies. Clin. Chem. 2007;53:1977-1979.

Chatterjee S, Lyle N, Mandal A, Kundu S. GSTT1 and GSTM1 gene deletions are not associated with hepatotoxicity caused by antitubercular drugs. J Clin Pharm Ther. 2010;35:465–470.

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

Chen B, Hu Y, Jin T, Lu D, Shao M, Zheng L, Wang Q, Shen Y, Liu H, Liu Y, Zhou Y. The influence of metabolic gene polymorphisms on urinary 1-hydroxypyrene concentrations in Chinese coke oven workers. Sci Total Environ. 2007;381(1-3):38-46.

Chen GF, Tang YM, Green B, Lin DX, Guengerich FP, Daly AK, Caporaso NE, Kadlubar FF. Low frequency of CYP2A6 gene polymorphism as revealed by a one-step polymerase chain reaction method. Pharmacogenetics. 1999;9(3):327-32.

Chida M, Yokoi T, Kosaka Y, Chiba K, Nakamura H, Ishizaki T, Yokota J, Kinoshita M, Sato K, Inaba M, Aoki Y, Gonzalez FJ, Kamataki T. Genetic polymorphism of CYP2D6 in the Japanese population. Pharmacogenetics. 1999;9(5):601-5.

Cho HJ, Lee SY, Ki CS, Kim JW. GSTM1, GSTT1 and GSTP1 polymorphisms in the Korean population. J Korean Med Sci. 2005;20(6):1089-92.

Christiansen L, Bygum A, Jensen A, Thomsen K, Brandrup F, Hørder M, et al. Association between CYP1A2 polymorphism and susceptibility to porphyria cutanea tarda. Hum Genet. 2000;107(6):612-4.

Chokunonga E, Borok MZ, Chirenje ZM, Nyakabau AM, Parkin DM. Trends in the incidence of cancer in the black population of Harare, Zimbabwe 1991-2010. Int J Cancer. 2013 Aug 1;133(3):721-9.

Coia V, Caglià A, Arredi B, Donati F, Santos FR, Pandya A, Taglioli L, Paoli G, Pascali V, Spedini G, Destro-Bisol G, Tyler-Smith C. Binary and microsatellite polymorphisms of the Y-chromosome in the Mbenzele pygmies from the Central African Republic. Am J Hum Biol. 2004;16(1):57-67.

Cornelis MC, El-Sohemy A, Campos H. Genetic polymorphism of CYP1A2 increases the risk of myocardial infarction. J Med Genet. 2004;41(10):758-62.

Correia C, Santos P, Coutinho AM, Vicente AM. Characterization of pharmacogenetically relevant CYP2D6 and ABCB1 gene polymorphisms in a Portuguese population sample. Cell Biochem Funct. 2009;27(4):251-5.

Coutinho P, Sandim V, Oliveira JA, Alves G, Hatagima A. Lack of association between glutathione S-transferase polymorphisms and primary glioma in a case-control study in Rio de Janeiro. Genet Mol Res. 2010 Mar 23;9(1):539-44.

Coutts RT. Polymorphism in the metabolism of drugs, including antidepressant drugs: comments on phenotyping. J Psychiatry Neurosci. 1994;19(1):30-44.

Coutts RT, Urichuk LJ. Polymorphic cytochromes P450 and drugs used in psychiatry. Cell Mol Neurobiol. 1999;19(3):325-54.

Crawhall N. Languages, genetics and archaeology: problems and the possibilties in Africa. In: Soodyall H (ed) The prehistory of Africa. Jonathan Ball Publishers, Johannesburg & Cape Town, 2006. pp 109-124.

Czerwinski M, Gibbs JP, Slattery JT. Busulfan conjugation by glutathione S-transferases alpha, mu, and pi. Drug Metab Dispos. 1996;24(9):1015-9.

da Silva J, Moraes CR, Heuser VD, Andrade VM, Silva FR, Kvitko K, et al. Evaluation of genetic damage in a Brazilian population occupationally exposed to pesticides and its correlation with polymorphisms in metabolizing genes. Mutagenesis. 2008;23(5):415-22.

Dalal S, Beunza JJ, Volmink J, Adebamowo C, Bajunirwe F, Njelekela M, et al. Non-communicable diseases in sub-Saharan Africa: what we know now. Int J Epidemiol. 2011;40(4):885-901.

D'Alò F, Voso MT, Guidi F, Massini G, Scardocci A, Sica S, Pagano L, Hohaus S, Leone G. Polymorphisms of CYP1A1 and glutathione S-transferase and susceptibility to adult acute myeloid leukemia. Haematologica. 2004;89(6):664-70.

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

Dandara C, Sayi J, Masimirembwa CM, Magimba A, Kaaya S, De Sommers K, et al. Genetic polymorphism of cytochrome P450 1A1 (Cyp1A1) and glutathione transferases (M1, T1 and P1) among Africans. Clin Chem Lab Med. 2002;40:952–957.

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

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

Dandara C, Li DP, Walther G, Parker MI.Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. Carcinogenesis. 2006;27(4):791-7.

Dandara C, Lombard Z, Du Plooy I, McLellan T, Norris SA, Ramsay M. Genetic variants in CYP (-1A2, -2C9, -2C19, -3A4 and -3A5), VKORC1 and ABCB1 genes in a black South African population: a window into diversity. Pharmacogenomics. 2011;12(12):1663-70.

Darazy M, Balbaa M, Mugharbil A, Saeed H, Sidani H, Abdel-Razzak Z. CYP1A1, CYP2E1, and GSTM1 gene polymorphisms and susceptibility to colorectal and gastric cancer among Lebanese. Genet Test Mol Biomarkers. 2011;15(6):423-9.

Deloménie C, Sica L, Grant DM, Krishnamoorthy R, Dupret JM. Genotyping of the polymorphic N-acetyltransferase (NAT2*) gene locus in two native African populations. Pharmacogenetics. 1996;6(2):177-85.

Deng SQ, Zeng XT, Wang Y, Ke Q, Xu QL. Meta-analysis of the CYP1A2 -163C>A Polymorphism and Lung Cancer Risk. Asian Pac J Cancer Prev. 2013;14(5):3155-8.

Dietzsch E, Laubscher R, Parker MI. Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. Int J Cancer. 2003;107(1):38-45.

Djordjevic N, Carrillo JA, van den Broek MP, Kishikawa J, Roh HK, Bertilsson L, Aklillu E. Comparisons of CYP2A6 genotype and enzyme activity between Swedes and Koreans. Drug Metab Pharmacokinet. 2012.

Dura P, Salomon J, Te Morsche RH, Roelofs HM, Kristinsson JO, Wobbes T, Witteman BJ, Tan AC, Drenth JP, Peters WH. No role for glutathione S-transferase genotypes in Caucasian esophageal squamous cell or adenocarcinoma etiology: an European case-control study. BMC Gastroenterol. 2013 Jun 3;13:97.

Ehret C. Languages and history. In: Heine and Nurse, editors. African Languages: An Introduction. Cambridge University Press, Cambridge, 2000. pp 272-297.

El-Shennawy GA, Elbialy AA, Isamil AE, El Behery MM. Is genetic polymorphism of ER-α, CYP1A1, and CYP1B1 a risk factor for uterine leiomyoma? Arch Gynecol Obstet. 2011;283(6):1313-8.

Esteller M, García A, Martínez-Palones JM, Xercavins J, Reventós J. Susceptibility to endometrial cancer: influence of allelism at p53, glutathione S-transferase (GSTM1 and GSTT1) and cytochrome P-450 (CYP1A1) loci. Br J Cancer. 1997;75(9):1385-8.

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

Excoffier, L. G. Laval, and S. Schneider. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online. 2005;1:47-50.

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567-87.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010;127(12):2893-917.

Fernandez-Salguero P, Hoffman SM, Cholerton S, Mohrenweiser H, Raunio H, Rautio A, et al. A genetic polymorphism in coumarin 7-hydroxylation: sequence of the human CYP2A genes and identification of variant CYP2A6 alleles. Am J Hum Genet. 1995 Sep;57(3):651-60.

Ferreira PM, Medeiros R, Vasconcelos A, Costa S, Pinto D, Morais A, et al. Association between CYP2E1 polymorphisms and susceptibility to prostate cancer. Eur J Cancer Prev. 2003;12(3):205-11.

Fischer K, Da Costa LA, García-Bailo B, Borchers CH, El-Sohemy A. Glutathione S-transferase (GST) M1, but not GSTT1, genotype influences plasma proteomic profiles in Caucasian and East Asian young adults. J Proteome Res. 2012 Oct 5;11(10):5022-33.

Flockhart DA (2007). "Drug Interactions: Cytochrome P450 Drug Interaction Table". Indiana University School of Medicine. http://medicine.iupui.edu/flockhart/table.htm. Retrieved on July 2011.

Fohner A, Muzquiz LI, Austin MA, Gaedigk A, Gordon A, Thornton T, Rieder MJ, Pershouse MA, Putnam EA, Howlett K, Beatty P, Thummel KE, Woodahl EL.

Pharmacogenetics in American Indian populations: analysis of CYP2D6, CYP3A4, CYP3A5, and CYP2C9 in the Confederated Salish and Kootenai Tribes. Pharmacogenet Genomics. 2013;23(8):403-414.

Forsberg LA, Absher D, Dumanski JP. Republished: Non-heritable genetics of human disease: spotlight on post-zygotic genetic variation acquired during lifetime. Postgrad Med J. 2013;89(1053):417-26.

Fransen M, Nordgren M, Wang B, Apanasets O, Van Veldhoven PP. Aging, age-related diseases and peroxisomes. Subcell Biochem. 2013;69:45-65.

Fujihara J, Shiwaku K, Xue Y, Kataoka K, Hieda Y, Takeshita H. CYP1A2 polymorphism (C > A at position -163) in Ovambos, Koreans and Mongolians. Cell Biochem Funct. 2007;25(5):491-4.

Fujihara J, Yasuda T, Iida R, Takatsuka H, Fujii Y, Takeshita H. Cytochrome P450 1A1, glutathione S-transferases M1 and T1 polymorphisms in Ovambos and Mongolians. Leg Med (Tokyo). 2009;11 Suppl 1:S408-10.

Furberg AH and CB Ambrosone. Molecular epidemiology, biomarkers and cancer prevention. Trends Mol Med. 2001;7(11)517-521.

Fuselli S, Gilman RH, Chanock SJ, Bonatto SL, De Stefano G, Evans CA, Labuda D, Luiselli D, Salzano FM, Soto G, Vallejo G, Sajantila A, Pettener D, Tarazona-Santos E. Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. Pharmacogenomics J. 2007;7:144-152.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177-83.

Gaikovitch EA, Cascorbi I, Mrozikiewicz PM, Brockmoller J, Frotschl R, Kopke K, Gerloff T, Chernov JN, Roots I. Polymorphisms of drug-metabolizing enzymes CYP2C9, CYP2C19, CYP2D6, CYP1A1, NAT2 and of P-glycoprotein in a Russian population. Eur. J. Clin. Pharmacol. 2003;59:303-312.

Gajecka M, Rydzanicz M, Jaskula-Sztul R, Kujawski M, Szyfter W, Szyfter K. CYP1A1, CYP2D6, CYP2E1, NAT2, GSTM1 and GSTT1 polymorphisms or their combinations are associated with the increased risk of the laryngeal squamous cell carcinoma. Mutat Res. 2005;574(1-2):112-23.

Gallegos-Arreola MP, Batista-González CM, Delgado-Lamas JL, Figuera LE, Puebla-Pérez AM, Arnaud-López L, Peralta-Leal V, Ramírez-Jirano LJ, Zúñiga-González GM. Cytochrome P4501A1 polymorphism is associated with susceptibility to acute lymphoblastic leukemia in adult Mexican patients. Blood Cells Mol Dis. 2004;33(3):326-9.

Galván CA, Elbarcha OC, Fernández EJ, Beltramo DM, Soria NW. Genetic profiling of GSTP1, DPYD, FCGR2A, FCGR3A and CCND1 genes in an Argentinian population. Clin Biochem. 2011;44(13):1058-61.

Garte S. The role of ethnicity in cancer susceptibility gene polymorphisms: the example of CYP1A1. Carcinogenesis. 1998;19(8):1329-32.

Gaspar PA, Hutz MH, Salzano FM, Hill K, Hurtado AM, Petzl-Erler ML, et al. Polymorphisms of CYP1a1, CYP2e1, GSTM1, GSTT1, and TP53 genes in Amerindians. Am J Phys Anthropol. 2002;119:249–256.

Gatta´s GJ, Kato M, Soares-Vieira JA, Siraque MS, Kohler P, Gomes L, et al. Ethnicity and glutathione S-transferase (GSTM1/GSTT1) polymorphisms in a Brazilian population. Braz J Med Biol Res. 2004;37:451–458.

Gawronska-Szklarz B, Wójcicki M, Kuprianowicz A, Kedzierska K, Kedzierski M, Górnik W, Pawlik A. CYP2D6 and GSTM1 genotypes in a Polish population. Eur J Clin Pharmacol. 1999;55(5):389-92.

Gemignani F, Landi S, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fabianova E, Mates D, Foretova L, Janout V, Bencko V, Gaborieau V, Gioia-Patricola L, Bellini I, Barale R, Canzian F, Hall J, Boffetta P, Hung RJ, Brennan P. Development of lung cancer before the age of 50: the role of xenobiotic metabolizing genes. Carcinogenesis. 2007;28(6):1287-93.

Glaubitz J.C. CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. Molecular Ecology Notes. 2004;4: 309-310.

Goldstein DB. Pharmacogenetics in the laboratory and the clinic. N Engl J Med. 2003;348(6):553-6.

Gong M, Dong W, Shi Z, Xu Y, Ni W, An R. Genetic polymorphisms of GSTM1, GSTT1, and GSTP1 with prostate cancer risk: a meta-analysis of 57 studies. PLoS One. 2012;7(11):e50587.

González MV, Alvarez V, Pello MF, Menéndez MJ, Suárez C, Coto E. Genetic polymorphism of N-acetyltransferase-2, glutathione S-transferase-M1, and cytochromes P450IIE1 and P450IID6 in the susceptibility to head and neck cancer. J Clin Pathol. 1998;51(4):294-8.

Gordillo-Bastidas E, Panduro A, Gordillo-Bastidas D, Zepeda-Carrillo EA, García-Bañuelos JJ, Muñoz-Valle JF, et al. Polymorphisms of alcohol metabolizing enzymes in indigenous Mexican population: unusual high frequency of CYP2E1*c2 allele. Alcohol Clin Exp Res. 2010;34(1):142-9.

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

Grimes BF (2001). Ethnologue: languages of the world. 13th ed. Summer Institute of Linguistics, Dallas.

Gronau S, Koenig-Greger D, Jerg M, Riechelmann H. Gene polymorphisms in detoxification enzymes as susceptibility factor for head and neck cancer? Otolaryngol Head Neck Surg. 2003;128(5):674-80.

Gross M, Kruisselbrink T, Anderson K, Lang N, McGovern P, Delongchamp R, Kadlubar F. Distribution and concordance of N-acetyltransferase genotype and phenotype in an American population. Cancer Epidemiol Biomarkers Prev. 1999 Aug;8(8):683-92.

Güldemann T, Vossen R. Khoisan. In: Heine B, Nurse D, editors. African languages. Cambridge: Cambridge University Press, 2002. p99–122.

Güldemann T, Stoneking M. A historical appraisal of clicks: A linguistic and genetic population perspective. Annu Rev Anthropol 2008;37:93–109.

Gundacker C, Komarnicki G, Jagiello P, Gencikova A, Dahmen N, Wittmann KJ, et al. Glutathione-S-transferase polymorphism, metallothionein expression, and mercury levels among students in Austria. Sci Total Environ. 2007;385:37–47.

Guthrie M (1948). The classification of Bantu languages. London: IAI/OUP.

Gutman G, Morad T, Peleg B, Peretz C, Bar-Am A, Safra T, et al. CYP1A1 and CYP2D6 gene polymorphisms in Israeli Jewish women with cervical cancer. Int J Gynecol Cancer. 2009;19(8):1300-2.

Gutman G, Morad T, Peleg B, Peretz C, Bar-Am A, Safra T, Grisaru D. CYP1A1 and CYP2D6 gene polymorphisms in Israeli Jewish women with cervical cancer. Int J Gynecol Cancer. 2009 Nov;19(8):1300-2.

Hammer O, Harper DAT and Ryan PD. PAST: Palaeontological Statistics software package for education and data analysis. Palaeontologia Electronica. 2000;4: 9.

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

Hamdy SI, Hiratsuka M, Narahara K, Endo N, El-Enany M, Moursi N, et al. Genotype and allele frequencies of TPMT, NAT2, GST, SULT1A1 and MDR-1 in the Egyptian population. Br J Clin Pharmacol. 2003a;55:560–569.

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

Han XM, Chen XP, Wu QN, Jiang CH, Zhou HH. G-2964A and C734A genetic polymorphisms of CYP1A2 in Chinese population. Acta Pharmacol Sin. 2000;21(11):1031-4.

Han S, Choi S, Chun YJ, Yun CH, Lee CH, Shin HJ, Na HS, Chung MW, Kim D. Functional characterization of allelic variants of polymorphic human cytochrome P450 2A6 (CYP2A6*5, *7, *8, *18, *19, and *35). Biol Pharm Bull. 2012;35(3):394-9.

Hardy GH. Mendelian Proportions in a Mixed Population. Science. 1908 Jul 10;28(706):49-50.

Hashemi M, Eskandari-Nasab E, Fazaeli A, Taheri M, Rezaei H, Mashhadi M, Arbabi F, Kaykhaei MA, Jahantigh M, Bahari G. Association between polymorphisms of glutathione S-transferase genes (GSTM1, GSTP1 and GSTT1) and breast cancer risk in a sample Iranian population. Biomark Med. 2012;6(6):797-803.

Hashemi-Soteh SM, Sarzare F, Merat F, Salehifar E, Shiran MR. Frequencies of three CYP2D6 nonfunctional alleles (CYP2D6*3, *4, and *6) within an Iranian population (Mazandaran). Genet Test Mol Biomarkers. 2011;15(11):821-5.

Hegele RA, Kwan K, Harris SB, Hanley AJ, Zinman B, Cao H. NAT2 polymorphism associated with plasma glucose concentration in Canadian Oji-Cree. Pharmacogenetics. 2000;10:233-238.

Hein DW, Doll MA, Fretland AJ, Leff MA, Webb SJ, Xiao GH, et al. Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. Cancer Epidemiol Biomarkers Prev. 2000;9(1):29-42.

Hein DW. Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. Mutat Res. 2002;506-507:65-77.

Heine B and Nurse D. Introduction. In: Heine and Nurse editors. African Languages: An Introduction. Cambridge University Press, Cambridge, 2000. pp 1-10.

Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci U S A. 2008;105(31):10693-8.

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, Bustamante CD, Mountain JL, Feldman MW. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A. 2011;108(13):5154-62.

Hersberger M, Marti-Jaun J, Rentsch K, Hänseler E. Rapid detection of the CYP2D6*3, CYP2D6*4, and CYP2D6*6 alleles by tetra-primer PCR and of the CYP2D6*5 allele by multiplex long PCR. Clin Chem. 2000;46(8 Pt 1):1072-7.

Hildesheim A, Chen CJ, Caporaso NE, Cheng YJ, Hoover RN, Hsu MM, Levine PH, Chen IH, Chen JY, Yang CS, et al. Cytochrome P4502E1 genetic polymorphisms and risk of nasopharyngeal carcinoma: results from a case-control study conducted in Taiwan. Cancer Epidemiol Biomarkers Prev. 1995 Sep;4(6):607-10.

Hildesheim A, Anderson LM, Chen CJ, Cheng YJ, Brinton LA, Daly AK, et al. CYP2E1 genetic polymorphisms and risk of nasopharyngeal carcinoma in Taiwan. J Natl Cancer Inst. 1997 Aug 20;89(16):1207-12.

Hsieh HI, Chen PC, Wong RH, Wang JD, Yang PM, Cheng TJ. Effect of the CYP2E1 genotype on vinyl chloride monomer-induced liver fibrosis among polyvinyl chloride workers. Toxicology. 2007;239(1-2):34-44.

http://www.who.int/mediacentre/factsheets/fs104/en/

http://www.ethnologue.com/

Huang YS, Su WJ, Huang YH, Chen CY, Chang FY, Lin HC, Lee SD. Genetic polymorphisms of manganese superoxide dismutase, NAD(P)H:quinone

oxidoreductase, glutathione S-transferase M1 and T1, and the susceptibility to drug-induced liver injury. J Hepatol. 2007;47:128–134.

Huang YS, Su WJ, Huang YH. Genetic polymoprphism of manganese dismutase, NAD(P)H: quinone oxidoreductase, glutathione S-transferase M1 and T1, and susceptibility to drug-induced liver injury. J Hepatol. 2007;47(1):128-34.

Huo R, Tang K, Wei Z, Shen L, Xiong Y, Wu X, Niu J, Han X, Tian Z, Yang L, Feng G, He L, Qin S. Genetic polymorphisms in CYP2E1: association with schizophrenia susceptibility and risperidone response in the Chinese Han population. PLoS One. 2012;7(5):e34809.

Iarmarcovai G, Sari-Minodier I, Orsière T, De Méo M, Gallice P, Bideau C, Iniesta D, Pompili J, Bergé-Lefranc JL, Botta A. A combined analysis of XRCC1, XRCC3, GSTM1 and GSTT1 polymorphisms and centromere content of micronuclei in welders. Mutagenesis. 2006 Mar;21(2):159-65.

Ingelman-Sundberg M, Oscarson M, McLellan RA. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. Trends Pharmacol Sci. 1999;20(8):342-9.

Ingelman-Sundberg M, Daly AK, Nebert DW. Human cytochrome P450 (CYP) allele nomenclature. 2008;Available at www.imm.ki.se=CYPalleles.

International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.

Islam MS, Ahmed MU, Sayeed MS, Maruf AA, Mostofa AG, Hussain SM, Kabir Y, Daly AK, Hasnat A. Lung cancer risk in relation to nicotinic acetylcholine receptor, CYP2A6 and CYP1A1 genotypes in the Bangladeshi population. Clin Chim Acta. 2013;416:11-9. Epub 2012.

Jemal A, Bray F, Forman D, O'Brien M, Ferlay J, Center M, Parkin DM. Cancer burden in Africa and opportunities for prevention. Cancer. 2012 Sep 15;118(18):4372-84.

Jenkins T. The prehistory of the San and Khoikhoi as recorded in their blood. Contemporary Studies on Khoisan. Vossen R and Keuthmann K. Hamburg, Helmut Buske Verlag. 1986;2: 51-77.

Ji L, Pan S, Wu J, Marti-Jaun J, Hersberger M. Genetic polymorphisms of CYP2D6 in Chinese mainland. Chin Med J (Engl). 2002;115(12):1780-4.

Jiang Z, Dalton TP, Jin L, Wang B, Tsuneoka Y, Shertzer HG, Deka R, Nebert DW. Toward the evaluation of function in genetic variability: characterizing human SNP frequencies and establishing BAC-transgenic mice carrying the human CYP1A1_CYP1A2 locus. Hum Mutat. 2005 Feb;25(2):196-206.

Jorde LB, Bamshad M, Rogers AR. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. Bioessays.1998;20(2):126-36.

Jorge LF, Eichelbaum M, Griese EU, Inaba T, Arias TD. Comparative evolutionary pharmacogenetics of CYP2D6 in Ngawbe and Embera Amerindians of Panama and Colombia: role of selection versus drift in world populations. Pharmacogenetics. 1999;9(2):217-28.

Jorge-Nebert LF, Eichelbaum M, Griese EU, Inaba T, Arias TD. Analysis of six SNPs of NAT2 in Ngawbe and Embera Amerindians of Panama and determination of the Embera acetylation phenotype using caffeine. Pharmacogenetics. 2002;12(1):39-48.

Ketelslegers HB, Gottschalk RW, Godschalk RW, Knaapen AM, van Schooten FJ, Vlietinck RF, et al. Interindividual variations in DNA adduct levels assessed by analysis of multiple genetic polymorphisms in smokers. Cancer Epidemiol Biomarkers Prev. 2006;15(4):624-9.

Khvostova EP, Pustylnyak VO, Gulyaeva LF. Genetic polymorphism of estrogen metabolizing enzymes in Siberian women with breast cancer. Genet Test Mol Biomarkers. 2012;16(3):167-73.

Kim HN, Kim NY, Yu L, Tran HT, Kim YK, Lee IK, Shin MH, Park KS, Choi JS, Kim HJ.Association of GSTT1 polymorphism with acute myeloid leukemia risk is dependent on smoking status. Leuk Lymphoma. 2012;53(4):681-7.

Kim JS, Park JY, Chung WY, Choi MA, Cho KS, Park KK. Polymorphisms in genes coding for enzymes metabolizing smoking-derived substances and the risk of periodontitis. J Clin Periodontol. 2004;31(11):959-64.

Kim JW, Lee CG, Park YG, Kim KS, Kim IK, Sohn YW, Min HK, Lee JM, Namkoong SE. Combined analysis of germline polymorphisms of p53, GSTM1, GSTT1, CYP1A1, and CYP2E1: relation to the incidence rate of cervical carcinoma. Cancer. 2000;88(9):2082-91.

Kim RB, Yamazaki H, Chiba K, O'Shea D, Mimura M, Guengerich FP, et al. In vivo and in vitro characterization of CYP2E1 activity in Japanese and Caucasians. J Pharmacol Exp Ther. 1996;279(1):4-11.

Kim SH, Kim SH, Yoon HJ, Shin DH, Park SS, Kim YS, Park JS, Jee YK. GSTT1 and GSTM1 null mutations and adverse reactions induced by antituberculosis drugs in Koreans. Tuberculosis (Edinb). 2010;90:39–43.

Kiran M, Chawla YK, Kaur J. Glutathione-S-transferase and microsomal epoxide hydrolase polymorphism and viral-related hepatocellular carcinoma risk in India. DNA Cell Biol. 2008;27(12):687-94.

Kiruthiga PV, Kannan MR, Saraswathi C, Pandian SK, Devi KP. CYP1A1 gene polymorphisms: lack of association with breast cancer susceptibility in the southern region (Madurai) of India. Asian Pac J Cancer Prev. 2011;12(8):2133-8.

Kiyohara C, Horiuchi T, Takayama K, Nakanishi Y. Genetic polymorphisms involved in carcinogen metabolism and DNA repair and lung cancer risk in a Japanese population. J Thorac Oncol. 2012;7(6):954-62.

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL. African Y chromosome and mtDNA divergence

provides insight into the history of click languages. Curr Biol. 2003;13(6):464-73.

Kotsopoulos J, Ghadirian P, El-Sohemy A, Lynch HT, Snyder C, Daly M, Domchek S, Randall S, Karlan B, Zhang P, Zhang S, Sun P, Narod SA. The CYP1A2 genotype modifies the association between coffee consumption and breast cancer risk among BRCA1 mutation carriers. Cancer Epidemiol Biomarkers Prev. 2007;16(5):912-6.

Kouhi H, Hamzeiy H, Barar J, Asadi M, Omidi Y. Frequency of five important CYP2D6 alleles within an Iranian population (Eastern Azerbaijan). Genet Test Mol Biomarkers. 2009;13(5):665-70.

Krajinovic M, Labuda D, Richer C, Karimi S, Sinnett D. Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. Blood. 1999;93(5):1496-501.

Kurahashi N, Sata F, Kasai S, Shibata T, Moriya K, Yamada H, et al. Maternal genetic polymorphisms in CYP1A1, GSTM1 and GSTT1 and the risk of hypospadias. Mol Hum Reprod. 2005;11(2):93-8.

Kurose K, Sugiyama E, Saito Y. Population differences in major functional polymorphisms of pharmacokinetics/pharmacodynamics-related genes in Eastern Asians and Europeans: implications in the clinical trials for novel drug development. Drug Metab Pharmacokinet. 2012;27(1):9-54.

Lakhdar R, Denden S, Knani J, Leban N, Daimi H, Hassine M, et al. Relationship between glutathione S-transferase P1 polymorphisms and chronic obstructive pulmonary disease in a Tunisian population. Genet Mol Res. 2010;9(2):897-907.

Lambert CA, Tishkoff SA. Genetic structure in African populations: implications for human demographic history. Cold Spring Harb Symp Quant Biol. 2009;74:395-402.

Lammers LA, Mathijssen RH, van Gelder T, Bijl MJ, de Graan AJ, Seynaeve C, van Fessem MA, Berns EM, Vulto AG, van Schaik RH. The impact of CYP2D6-predicted phenotype on tamoxifen treatment outcome in patients with metastatic breast cancer. Br J Cancer. 2010 Sep 7;103(6):765-71.

Landi MT, Bertazzi PA, Shields PG, Clark G, Lucier GW, Garte SJ, et al. Association between CYP1A1 genotype, mRNA expression and enzymatic activity in humans. Pharmacogenetics. 1994;4(5):242-6.

Landi S, Gemignani F, Moreno V, Gioia-Patricola L, Chabrier A, Guino E, Navarro M, de Oca J, Capellà G, Canzian F; Bellvitge Colorectal Cancer Study Group. A comprehensive analysis of phase I and phase II metabolism gene polymorphisms and risk of colorectal cancer. Pharmacogenet Genomics. 2005;15(8):535-46.

Lavandera JV, Parera VE, Batlle A, Buzaleh AM. CYP2D6 polymorphisms in patients with porphyrias. Mol Med. 2006;12(9-10):259-63.

Leathart JB, London SJ, Steward A, Adams JD, Idle JR, Daly AK. CYP2D6 phenotype-genotype relationships in African-Americans and Caucasians in Los Angeles. Pharmacogenetics. 1998;8(6):529-41.

Lee RB (1979). The !Kung San: men, women, and work in a foraging society. Cambridge, Cambridge University Press.

Lee SY, Lee KA, Ki CS, Kwon OJ, Kim HJ, Chung MP, Suh GY, Kim JW. Complete sequencing of a genetic polymorphism in NAT2 in the Korean population. Clin. Chem. 2002;48:775-777.

Leichsenring A, Losi-Guembarovski R, Maciel ME, Losi-Guembarovski A, Oliveira BW, Ramos G, et al. CYP1A1 and GSTP1 polymorphisms in an oral cancer case-control study. Braz J Med Biol Res. 2006;39(12):1569-74.

Leiro V, Ferna´ndez-Villar A, Valverde D, Constenla L, Va´zquez R, Pin˜ eiro L, Gonza´lez-Quintela A. Influence of glutathione S-transferase M1 and T1

homozygous null mutations on the risk of antituberculosis drug-induced hepatotoxicity in a Caucasian population. Liver Int. 2008;28:835–839.

Leiro V, Fernandez-Villar A, Valverde D, Constenla L, Vazquez R, Pineiro L, et al. Influence of glutathione S-transferase M1 and T1 homozygous null mutations on the risk of antituberculosis drug–induced hepatotoxicity in a Caucasian population. Liver Int. 2008;28(6):835-9.

Li D, Dandara C, Parker MI. Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. Clin Chem Lab Med. 2005;43(4):370-5.

Li D, Dandara C, Parker MI. The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. BMC Genet. 2010;11:47.

Li DP, Dandara C, Walther G, Parker MI. Genetic polymorphisms of alcohol metabolising enzymes: their role in susceptibility to oesophageal cancer. Clin Chem Lab Med. 2008;46(3):323-8.

Liang S, Wei X, Gong C, Wei J, Chen Z, Chen X, Wang Z, Deng J. Significant association between asthma risk and the GSTM1 and GSTT1 deletion polymorphisms: An updated meta-analysis of case-control studies. Respirology. 2013;18(5):774-83.

Lin HJ, Han CY, Lin BK, Hardy S. Ethnic distribution of slow acetylator mutations in the polymorphic N-acetyltransferase (NAT2) gene. Pharmacogenetics. 1994;4(3):125-34.

Linder MW, Prough RA, Valdes R Jr. Pharmacogenetics: a laboratory tool for optimizing therapeutic efficiency. Clin Chem. 1997;43(2):254-66.

Liu L, Li C, Gao J, Li K, Gao L, Gao T. Genetic polymorphisms of glutathione S-transferase and risk of vitiligo in the Chinese population. J Invest Dermatol. 2009;129:2646–2652.

Loktionov A, Moore W, Spencer SP, Vorster H, Nell T, O'Neill IK, et al. Differences in N-acetylation genotypes between Caucasians and Black South Africans: implications for cancer prevention. Cancer Detect Prev. 2002;26(1):15-22.

London SJ, Idle JR, Daly AK, Coetzee GA. Genetic variation of CYP2A6, smoking, and risk of cancer. Lancet. 1999;353(9156):898-9.

López-Cima MF, Alvarez-Avellón SM, Pascual T, Fernández-Somoano A, Tardón A. Genetic polymorphisms in CYP1A1, GSTM1, GSTP1 and GSTT1 metabolic genes and risk of lung cancer in Asturias. BMC Cancer. 2012;12:433.

Losi-Guembarovski R, Cólus IM, De Menezes RP, Poliseli F, Chaves VN, Kuasne H, et al. Lack of association among polymorphic xenobiotic-metabolizing enzyme genotypes and the occurrence and progression of oral carcinoma in a Brazilian population. Anticancer Res. 2008;28(2A):1023-8.

Loyha K, Vatanasapt P, Promthet S, Parkin DM. Risk factors for oral cancer in northeast Thailand. Asian Pac J Cancer Prev. 2012;13(10):5087-90.

Lu SH, Ohshima H, Fu HM, Tian Y, Li FM, Blettner M, Wahrendorf J, Bartsch H. Urinary excretion of N-nitrosamino acids and nitrate by inhabitants of high- and low-risk areas for esophageal cancer in Northern China: endogenous formation of nitrosoproline and its inhibition by vitamin C. Cancer Res. 1986 Mar;46(3):1485-91.

Luca F, Bubba G, Basile M, Brdicka R, Michalodimitrakis E, Rickards O, Vershubsky G, Quintana-Murci L, Kozlov AI, Novelletto A. Multiple advantageous amino acid variants in the NAT2 gene in human populations. PLoS ONE 2008;3: e3136.

Lucena MI, Andrade RJ, Martínez C, Ulzurrun E, García-Martín E, Borraz Y, Fernández MC, Romero-Gomez M, Castiella A, Planas R, Costa J, Anzola S, Agúndez JA; Spanish Group for the Study of Drug-Induced Liver Disease. Glutathione S-transferase m1 and t1 null genotypes increase susceptibility to idiosyncratic drug-induced liver injury. Hepatology. 2008;48:588–596.

MacEachern S. Genes, Tribes, and African History. Curr Anthropol. 2000;41(3):357-384.

Magno LA, Talbot J, Talbot T, Borges Santos AM, Souza RP, Marin LJ, et al. Glutathione S-Transferase variants in a Brazilian population. Pharmacology. 2009;83:231–236.

Martinez C, Agundez JA, Olivera M, Llerena A, Ramirez R, Hernandez M, Benitez J. Influence of genetic admixture on polymorphisms of drug-metabolizing enzymes: analyses of mutations on NAT2 and C gamma P2E1 genes in a mixed Hispanic population. Clin. Pharmacol. Ther. 1998;63:623-628.

Martinez-Labarga C, Lelli R, Tarsi T, Babalini C, De Angelis F, Ottoni C, et al. Polymorphisms of the COL1A2, CYP1A1 and HS1,2 Ig enhancer genes in the Tuaregs from Libya. Ann Hum Biol. 2007;34(4):425-36.

Martínez-Ramírez OC, Pérez-Morales R, Castro C, Flores-Díaz A, Soto-Cruz KE, Astorga-Ramos A, Gonsebatt ME, Casas L, Valdés-Flores M, Rubio J. Polymorphisms of catechol estrogens metabolism pathway genes and breast cancer risk in Mexican women. Breast. 2012; pii: S0960-9776(12)00172-5.

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

Matejcic M, Li D, Prescott NJ, Lewis CM, Mathew CG, Parker MI. Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. PLoS One. 2011;6(12):e29366.

McLellan RA, Oscarson M, Seidegård J, Evans DA, Ingelman-Sundberg M. Frequent occurrence of CYP2D6 gene duplication in Saudi Arabians. Pharmacogenetics. 1997;7(3):187-91.

Menoyo A, del Rio E, Baiget M. Characterization of variant alleles of cytochrome CYP2D6 in a Spanish population. Cell Biochem Funct. 2006;24(5):381-5.

Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1998;16:1215.

Millikan R, Pittman G, Tse CK, Savitz DA, Newman B, Bell D.Glutathione S-transferases M1, T1, and P1 and breast cancer. Cancer Epidemiol Biomarkers Prev. 2000;9:567–573.

Mirvish SS, Grandjean AC, Reimers KJ, Connelly BJ, Chen SC, Gallagher J, et al. Dosing time with ascorbic acid and nitrate, gum and tobacco chewing, fasting, and other factors affecting N-nitrosoproline formation in healthy subjects taking proline with a standard meal. Cancer Epidemiol Biomarkers Prev. 1995;4(7):775-82.

Mittal RD, Srivastava DS. Genetic polymorphism of drug metabolizing enzymes (CYP2E1, GSTP1) and susceptibility to bladder cancer in North India. Asian Pac J Cancer Prev. 2005;6(1):6-9.

Montano V, Marcari V, Pavanello M, Anyaele O, Comas D, Destro-Bisol G, Batini C. The influence of habitats on female mobility in Central and Western Africa inferred from human mitochondrial variation. BMC Evol Biol. 2013 Jan 29;13:24.

Monteiro TP, El-Jaick KB, Jeovanio-Silva AL, Brasil PE, Costa MJ, Rolla VC, de Castro L. The roles of GSTM1 and GSTT1 null genotypes and other predictors in anti-tuberculosis drug-induced liver injury. J Clin Pharm Ther. 2012;37(6):712-8.

Montero R, Araujo A, Carranza P, Mejı´a-Loza V, Serrano L, Albores A, et al. Genotype frequencies of polymorphic GSTM1, GSTT1, and cytochrome P450 CYP1A1 in Mexicans. Hum Biol. 2007;79:299–312.

Mrozikiewicz PM, Cascorbi I, Brockmoller J, Roots I. Determination and allelic allocation of seven nucleotide transitions within the arylamine N-

acetyltransferase gene in the Polish population. Clin. Pharmacol. Ther. 1996;59:376-382.

Murata M, Watanabe M, Yamanaka M, Kubota Y, Ito H, Nagao M, et al. Genetic polymorphisms in cytochrome P450 (CYP) 1A1, CYP1A2, CYP2E1, glutathione S-transferase (GST) M1 and GSTT1 and susceptibility to prostate cancer in the Japanese population. Cancer Lett. 2001;165(2):171-7.

Mwenifumbo JC, Myers MG, Wall TL, Lin SK, Sellers EM, Tyndale RF. Ethnic variation in CYP2A6*7, CYP2A6*8 and CYP2A6*10 as assessed with a novel haplotyping method. Pharmacogenet Genomics. 2005;15(3):189-92.

Naidoo T, Schlebusch CM, Makkan H, Patel P, Mahabeer R, Erasmus JC, et al. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Investig Genet. 2010;1(1):6.

Nakajima M, Yokoi T, Mizutani M, Kinoshita M, Funayama M, Kamataki T. Genetic polymorphism in the 5'-flanking region of human CYP1A2 gene: effect on the CYP1A2 inducibility in humans. J Biochem (Tokyo). 1999;125(4):803-8.

Nakajima M, Fukami T, Yamanaka H, Higashi E, Sakai H, Yoshida R, Kwon JT, McLeod HL, Yokoi T.. Comprehensive evaluation of variability in nicotine metabolism and CYP2A6 polymorphic alleles in four ethnic populations. Clin Pharmacol Ther. 2006 Sep;80(3):282-97.

National Cancer Registry, National Health Laboratory Service. 2006, http://www.nioh.ac.za/?page=national_cancer_registry&id=41

Naveen AT, Adithan C, Soya SS, Gerard N, Krishnamoorthy R. CYP2D6 genetic polymorphism in South Indian populations. Biol Pharm Bull. 2006;29(8):1655-8.

Nebert DW. Polymorphisms in drug-metabolizing enzymes: what is their clinical relevance and why do they exist? Am J Hum Genet. 1997;60(2):265-71.

Niewoehner WA. Behavioral inferences from the Skhul/Qafzeh early modern human hand remains. Proc Natl Acad Sci U S A 2001;98:2979–2984.

Nishino K, Sekine M, Kodama S, Sudo N, Aoki Y, Seki N, Tanaka K. Cigarette smoking and glutathione S-transferase M1 polymorphism associated with risk for uterine cervical cancer. J Obstet Gynaecol Res. 2008;34(6):994-1001.

Nurfadhlina M, Foong K, Teh LK, Tan SC, Mohd Zaki S, Ismail R. CYP2A6 polymorphisms in Malays, Chinese and Indians. Xenobiotica. 2006;36(8):684-92.

Nurse GT, Weiner JS, Jenkins T (1985). The peoples of southern Africa and their affinities. Oxford: Clarendon Press.

Okobia M, Bunker C, Zmuda J, Kammerer C, Vogel V, Uche E, et al. Cytochrome P4501A1 genetic polymorphisms and breast cancer risk in Nigerian women. Breast Cancer Res Treat. 2005;94(3):285-93.

Olshan AF, Weissler MC, Watson MA, Bell DA. GSTM1, GSTT1, GSTP1, CYP1A1, and NAT1 polymorphisms, tobacco use, and the risk of head and neck cancer. Cancer Epidemiol Biomarkers Prev. 2000;9(2):185-91.

Oscarson M, Gullstén H, Rautio A, Bernal ML, Sinues B, Dahl ML, et al. Genotyping of human cytochrome P450 2A6 (CYP2A6), a nicotine C-oxidase. FEBS Lett. 1998;438(3):201-5.

Ouerhani S, Marrakchi R, Bouhaha R, Ben Slama MR, Sfaxi M, Ayed M, et al. The role of CYP2D6*4 variant in bladder cancer susceptibility in Tunisian patients. Bull Cancer. 2008;95(2):E1-4.

Ouerhani S, Tebourski F, Slama MR, Marrakchi R, Rabeh M, Hassine LB, et al. The role of glutathione transferases M1 and T1 in individual susceptibility to bladder cancer in a Tunisian population. Ann Hum Biol. 2006;33(5-6):529-35.

Patin E, Barreiro LB, Sabeti PC, Austerlitz F, Luca F, Sajantila A, Behar DM, Semino O, Sakuntabhai A, Guiso N, Gicquel B, McElreavey K, Harding RM, Heyer E, Quintana-Murci L: Deciphering the ancient and complex evolutionary history

of human arylamine N-acetyltransferase genes. Am. J. Hum. Genet. 2006a;78:423-436.

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

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert JM, Gessain A, Froment A, Bahuchet S, Heyer E, Quintana-Murci L. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet. 2009;5(4):e1000448.

Parkin DM. International variation. Oncogene. 2004;23(38):6329-40.

Parkin DM, Sitas F, Chirenje M, Stein L, Abratt R, Wabinga H. Part I: Cancer in Indigenous Africans--burden, distribution, and trends. Lancet Oncol. 2008;9(7):683-92.

Parkin DM, Nambooze S, Wabwire-Mangen F, Wabinga HR. Changing cancer incidence in Kampala, Uganda, 1991-2006. Int J Cancer. 2010;126(5):1187-95.

Parveen F, Faridi RM, Das V, Tripathi G, Agrawal S. Genetic association of phase I and phase II detoxification genes with recurrent miscarriages among North Indian women. Mol Hum Reprod. 2010;16(3):207-14.

Peamkrasatam S, Sriwatanakul K, Kiyotani K, Fujieda M, Yamazaki H, Kamataki T, Yoovathaworn K. In vivo evaluation of coumarin and nicotine as probe drugs to predict the metabolic capacity of CYP2A6 due to genetic polymorphism in Thais. Drug Metab Pharmacokinet. 2006;21(6):475-84.

Pérez-Morales R, Castro-Hernández C, Gonsebatt ME, Rubio J. Polymorphism of CYP1A1*2C, GSTM1*0, and GSTT1*0 in a Mexican Mestizo population: a similitude analysis. Hum Biol. 2008;80(4):457-65.

Pereira L, Gusmão L, Alves C, Amorim A, Prata MJ. Bantu and European Y-lineages in Sub-Saharan Africa. Ann Hum Genet. 2002;66(Pt 5-6):369-78.

Persson I, Johansson I, Bergling H, Dahl ML, Seidegård J, Rylander R, et al. Genetic polymorphism of cytochrome P4502E1 in a Swedish population. Relationship to incidence of lung cancer. FEBS Lett. 1993;319(3):207-11.

Pharoah PD, Dunning AM, Ponder BA, Easton DF. Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer. 2004;4(11):850-60.

Philip Goelet, Michael R. Knapp, Stephen Anderson. U.S. Patent No 5,888,819 (1999). Washington, DC: U.S. Patent and Trademark Office.

Piacentini S, Polimanti R, Porreca F, Martínez-Labarga C, De Stefano GF, Fuciarelli M. GSTT1 and GSTM1 gene polymorphisms in European and African populations. Mol Biol Rep. 2011;38(2):1225-30.

Pianezza ML, Sellers EM, Tyndale RF. Nicotine metabolism defect reduces smoking. Nature. 1998;393(6687):750.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945-59.

Rafiee L, Saadat I, Saadat M. Glutathione S-transferase genetic polymorphisms (GSTM1, GSTT1 and GSTO2) in three Iranian populations. Mol Biol Rep. 2010;37:155–158.

Raymond M. and F. Rousset. An exact test for population differentiation. Evolution. 1995;49:1280-1283.

Razmkhah F, Pazhakh V, Zaker F, Atashrazm F, Sheikhi M. Frequency of CYP1A1*2C polymorphism in patients with leukemia in the Iranian population. Lab Med. 2011;42: 220–223.

Rebbeck TR, Troxel AB, Walker AH, Panossian S, Gallagher S, Shatalova EG, et al. Pairwise combinations of estrogen metabolism genotypes in postmenopausal breast cancer etiology. Cancer Epidemiol Biomarkers Prev. 2007;16(3):444-50.

Rightmire GP. Out of Africa: Modern Human Origins Special Feature: Middle and later Pleistocene hominins in Africa and Southwest Asia. Proc Natl Acad Sci U S A 2009;106:16046–16050.

Roco A, Quiñones L, Agúndez JA, García-Martín E, Squicciarini V, Miranda C, Garay J, Farfán N, Saavedra I, Cáceres D, Ibarra C, Varela N. Frequencies of 23 functionally significant variant alleles related with metabolism of antineoplastic drugs in the chilean population: comparison with caucasian and asian populations. Front Genet. 2012;3:229.

Rodrigo L, Alvarez V, Rodriguez M, Pérez R, Alvarez R, Coto E. N-acetyltransferase-2, glutathione S-transferase M1, alcohol dehydrogenase, and cytochrome P450IIE1 genotypes in alcoholic liver cirrhosis: a case-control study. Scand J Gastroenterol. 1999;34(3):303-7.

Roses AD. Pharmacogenetics and the practice of medicine. Nature. 2000;405(6788):857-65

Rosenberg NA. Distruct: a program for the graphical display of population structure. Molecular Ecology Notes. 2004;4:137-138.

Rostami-Hodjegan A, Lennard MS, Woods HF, Tucker GT.Meta-analysis of studies of the CYP2D6 polymorphism in relation to lung cancer and Parkinson's disease. Pharmacogenetics. 1998;8(3):227-38.

Roy B, Chowdhury A, Kundu S, Santra A, Dey B, Chakraborty M, et al. Increased risk of antituberculosis drug-induced hepatotoxicity in individuals with glutathione S-transferase M1 ‘null’ mutation. J Gastroenterol Hepatol. 2001;16(9):1033-7.

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000;132:365-386.

Rudolph A, Hein R, Hoffmeister M, Försti A, Hemminki K, Risch A, Brenner H, Chang-Claude J. Copy number variations of GSTT1 and GSTM1, colorectal

cancer risk and possible effect modification of cigarette smoking and menopausal hormone therapy. Int J Cancer. 2012 Sep 1;131(5):E841-8.

Sabbagh A, Langaney A, Darlu P, Gérard N, Krishnamoorthy R, Poloni ES. Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet. 2008;9:21.

Sachse C, Brockmöller J, Bauer S, Roots I. Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. Am J Hum Genet. 1997 Feb;60(2):284-95.

Saebø M, Skjelbred CF, Brekke Li K, Bowitz Lothe IM, Hagen PC, Johnsen E, Tveit KM, Kure EH. CYP1A2 164 A-->C polymorphism, cigarette smoking, consumption of well-done red meat and risk of developing colorectal adenomas and carcinomas. Anticancer Res. 2008 Jul-Aug;28(4C):2289-95.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A. The making of the African mtDNA landscape. Am J Hum Genet. 2002;71(5):1082-111.

Salazar-Flores J, Torres-Reyes LA, Martínez-Cortés G, Rubi-Castellanos R, Sosa-Macías M, Muñoz-Valle JF, González-González C, Ramírez A, Román R, Méndez JL, Barrera A, Torres A, Medina R, Rangel-Villalobos H. Distribution of CYP2D6 and CYP2C19 polymorphisms associated with poor metabolizer phenotype in five Amerindian groups and western Mestizos from Mexico. Genet Test Mol Biomarkers. 2012;16(9):1098-104.

Salem AH, Yaqoob A, Ali M, Handu S, Fadel R, Abu-Hijleh M, Almawi W. Genetic polymorphism of the glutathione S-transferase M1 and T1 genes in three distinct Arab populations. Dis Markers. 2011;31(5):311-6.

Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, et al. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. 2006;27(9):1713-24.

Sangrajrang S, Sato Y, Sakamoto H, Ohnami S, Khuhaprema T, Yoshida T. Genetic polymorphisms in folate and alcohol metabolism and breast cancer risk: a case-control study in Thai women. Breast Cancer Res Treat. 2010;123(3):885-93.

Sangrajrang S, Sato Y, Sakamoto H, Ohnami S, Laird NM, Khuhaprema T, Brennan P, Boffetta P, Yoshida T. Genetic polymorphisms of estrogen metabolizing enzyme and breast cancer risk in Thai women. Int J Cancer. 2009 Aug 15;125(4):837-43.

Santovito A, Burgarello C, Cervella P, Delpero M. Polymorphisms of cytochrome P450 1A1, glutathione s-transferases M1 and T1 genes in Ouangolodougou (Northern Ivory Coast). Genet Mol Biol. 2010;33(3):434-7.

Sarmanova J, Benesova K, Gut I, Nedelcheva-Kristensen V, Tynkova L, Soucek P. Genetic polymorphisms of biotransformation enzymes in patients with Hodgkin's and non-Hodgkin's lymphomas. Hum Mol Genet. 2001;10(12):1265-73.

Scheinfeldt LB, Soi S, Tishkoff SA. Colloquium paper: working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. Proc Natl Acad Sci U S A. 2010;107 Suppl 2:8931-8.

Schlebusch CM. Genetic variation in Khoisan-speaking populations from southern Africa. University of the Witwatersrand. 2010a.

Schlebusch C. Issues raised by use of ethnic-group names in genome study. Nature 2010b;464:487; author reply 487.

Schlebusch CM, de Jongh M, Soodyall H. Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. J Hum Genet. 2011 Sep;56(9):623-30.

Schlebusch CM, Soodyall H, Jakobsson M. Genetic variation of 15 autosomal STR loci in various populations from southern Africa. Forensic Sci Int Genet. 2012;6(1):e20-1.

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, Steenkamp AW, Oostuisen H, Venter P, Gajewski J, Zhang Y, Pugh BF, Makova KD, Nekrutenko A, Mardis ER, Patterson N, Pringle TH, Chiaromonte F, Mullikin JC, Eichler EE, Hardison RC, Gibbs RA, Harkins TT, Hayes VM. Complete Khoisan and Bantu genomes from southern Africa. Nature. 2010;463(7283):943-7.

Schwartzbaum JA, Ahlbom A, Lönn S, Warholm M, Rannug A, Auvinen A, Christensen HC, Henriksson R, Johansen C, Lindholm C, Malmer B, Salminen T, Schoemaker MJ, Swerdlow AJ, Feychting M. An international case-control study of glutathione transferase and functionally related polymorphisms and risk of primary adult brain tumors. Cancer Epidemiol Biomarkers Prev. 2007;16(3):559-65.

Selvin S. Cytochrome P450 1A1 polymorphism and childhood leukemia:an analysis of matched pairs case-control genotype data. Cancer Epidemiol Biomarkers Prev. 2004;13: 1371–1374.

Seremak-Mrozikiewicz A, Drews K, Semczuk A, Jakowicki JA, Mrozikiewicz PM. CYP1A1 alleles in female genital cancers in the Polish population. Eur J Obstet Gynecol Reprod Biol. 2005;118(2):246-50.

Sheweita SA. Drug-metabolizing enzymes: mechanisms and functions. Curr Drug Metab. 2000;1(2):107-32.

Shimada N, Iwasaki M, Kasuga Y, Yokoyama S, Onuma H, Nishimura H, Kusama R, Hamada GS, Nishimoto IN, Iyeyasu H, Motola J Jr, Laginha FM, Kurahashi N, Tsugane S. Genetic polymorphisms in estrogen metabolism and breast cancer risk in case-control studies in Japanese, Japanese Brazilians and non-Japanese Brazilians. J Hum Genet. 2009;54(4):209-15.

Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. Eur J Hum Genet. 2011;19(1):84-8.

Singh AP, Pant MC, Ruwali M, Shah PP, Prasad R, Mathur N, Parmar D. Polymorphism in cytochrome P450 1A2 and their interaction with risk factors in determining risk of squamous cell lung carcinoma in men. Cancer Biomark. 2010-2011;8(6):351-9.

Skrętkowicz J, Barańska M, Kaczorowska A, Rychlik-Sych M. Genetic polymorphisms of CYP2D6 oxidation in patients with systemic lupus erythematosus. Arch Med Sci. 2011;7(5):864-9.

Silveira Vda S, Canalle R, Scrideli CA, Queiroz RG, Tone LG. Role of the CYP2D6, EPHX1, MPO, and NQO1 genes in the susceptibility to acute lymphoblastic leukemia in Brazilian children. Environ Mol Mutagen. 2010 Jan;51(1):48-56.

Smith CA, Gough AC, Leigh PN, Summers BA, Harding AE, Maraganore DM, Sturman SG, et al. Debrisoquine hydroxylase gene polymorphism and susceptibility to Parkinson's disease. Lancet. 1992;339(8806):1375-7.

Soerjomataram I, Lortet-Tieulent J, Parkin DM, Ferlay J, Mathers C, Forman D, Bray F. Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. Lancet. 2012 Nov 24;380(9856):1840-50.

Somdyala NI, Bradshaw D, Gelderblom WC, Parkin DM. Cancer incidence in a rural population of South Africa, 1998-2002. Int J Cancer. 2010;127(10):2420-9.

Song DK, Xing DL, Zhang LR, Li ZX, Liu J, Qiao BP. Association of NAT2, GSTM1, GSTT1, CYP2A6, and CYP2A13 gene polymorphisms with susceptibility and clinicopathologic characteristics of bladder cancer in Central China. Cancer Detect. Prev. 2009;32:416-423.

Soodyall H, Jenkins T. Mitochondrial DNA polymorphisms in Khoisan populations from southern Africa. Ann Hum Genet. 1992a;56(Pt 4):315-24.

Soodyall H, Jenkins T. Mitochondrial DNA studies in the South African Indian population. Gene Geogr. 1992b;6(3):127-37.

Soodyall H, Jenkins T. Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of the Dama, Herero and Ambo. Ann Hum Biol. 1993;20(5):477-85.

Soodyall H, Makkan H, Haycock P, Naidoo T. The genetic prehistory of the Khoe and San. Southern African Humanities. 2008;20(1):37-48.

Souiden Y, Mahdouani M, Chaieb K, Bakhrouf A, Mahdouani K. Lack of association of CYP1A1 polymorphism with prostate cancer susceptibility of Tunisian men. Genet Test Mol Biomarkers. 2012;16(7):661-6.

Soyama A, Saito Y, Hanioka N, Maekawa K, Komamura K, Kamakura S, Kitakaze M, Tomoike H, Ueno K, Goto Y, Kimura H, Katoh M, Sugai K, Saitoh O, Kawai M, Ohnuma T, Ohtsuki T, Suzuki C, Minami N, Kamatani N, Ozawa S, Sawada J. Single nucleotide polymorphisms and haplotypes of CYP1A2 in a Japanese population. Drug Metab Pharmacokinet. 2005 Feb;20(1):24-33.

Staehli Hodel EM, Csajka C, Ariey F, Guidi M, Kabanywanyi AM, Duong S, Decosterd LA, Olliaro P, Beck HP, Genton B. Effect of single nucleotide polymorphisms in cytochrome p450 isoenzyme and N-acetyltransferase 2 genes on the metabolism of artemisinin-based combination therapies in malaria patients from cambodia and Tanzania. Antimicrob Agents Chemother. 2013;57(2):950-8. Epub 2012.

Su MW, Tung KY, Liang PH, Tsai CH, Kuo NW, Lee YL. Gene-gene and gene-environmental interactions of childhood asthma: a multifactor dimension reduction approach. PLoS One. 2012;7(2):e30694.

Su C, Andrew A, Karagas MR, Borsuk ME. Using Bayesian networks to discover relations between genes, environment, and disease. BioData Min. 2013 Mar 21;6(1):6.

Sugawara T, Nomura E, Sagawa T, Sakuragi N, Fujimoto S. CYP1A1 polymorphism and risk of gynecological malignancy in Japan. Int J Gynecol Cancer. 2003;13(6):785-90.

Sun JX, Mullikin JC, Patterson N, Reich DE. Microsatellites are molecular clocks that support accurate inferences about history. Mol Biol Evol. 2009;26(5):1017-27.

Taioli E, Ford J, Trachman J, Li Y, Demopoulos R, Garte S. Lung cancer risk and CYP1A1 genotype in African Americans. Carcinogenesis. 1998;19(5):813-7.

Tajima F (1993). Measurement of DNA polymorphism. In: Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology, edited by Takahata, N. and Clark, A.G., Tokyo, Sunderland, MA: Japan Scientific Societies Press, Sinauer Associates, Inc., p.37-59.

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989a;123:585-595.

Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989b;123:597-601.

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007;24(8):1596-9.

Tan W, Chen GF, Xing DY, Song CY, Kadlubar FF, Lin DX. Frequency of CYP2A6 gene deletion and its relation to risk of lung and esophageal cancer in the Chinese population. Int J Cancer. 2001;95(2):96-101.

Teixeira RL, Miranda AB, Pacheco AG, Lopes MQ, Fonseca-Costa J, Rabahi MF, et al. Genetic profile of the arylamine N-acetyltransferase 2 coding gene among individuals from two different regions of Brazil. Mutat Res. 2007;624(1-2):31-40.

The International HapMap Project 2008. www.hapmap.org.

Thier R, Brüning T, Roos PH, Rihs HP, Golka K, Ko Y, Bolt HM. Markers of genetic susceptibility in human environmental hygiene and toxicology: The role of selected CYP, NAT and GST genes. Int J Hyg Environ Health. 2003;206:149-171.

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, et al. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol. 2007;24(10):2180-95.

Todesco L, Török M, Krähenbühl S, Wenk M. Determination of -3858G-->A and -164C-->A genetic polymorphisms of CYP1A2 in blood and saliva by rapid allelic discrimination: large difference in the prevalence of the -3858G-->A mutation between Caucasians and Asians. Eur J Clin Pharmacol. 2003;59(4):343-6.

Torkaman-Boutorabi A, Hoormand M, Naghdi N, Bakhshayesh M, Milanian I. Genotype and allele frequencies of N-acetyltransferase 2 and glutathione S-transferase in the Iranian population. Clin. Exp. Pharmacol. Physiol. 2007;34:1207-1211.

Torres J, Lopez L, Lazcano E, Camorlinga M, Flores L, Muñoz O. Trends in Helicobacter pylori infection and gastric cancer in Mexico. Cancer Epidemiol Biomarkers Prev. 2005 Aug;14(8):1874-7.

Torresan C, Oliveira MM, Torrezan GT, de Oliveira SF, Abuázar CS, Losi-Guembarovski R, et al. Genetic polymorphisms in oestrogen metabolic pathway and breast cancer: a positive association with combined CYP/GST genotypes. Clin Exp Med. 2008;8(2):65-71.

Touré A, Diop C, Cabral M, Fall M, Lhermitte M, Diouf A, Broly F, Allorge D. Study of NAT2 genetic polymorphism in West African subjects: example of an healthy non-smoker Senegalese population. Mol Biol Rep. 2012 Dec;39(12):10489-96.Tsai PC, Huang W, Lee Y, Chan SH, Guo YL. Genetic

polymorphisms in CYP1A1 and GSTM1 predispose humans to PCBs/PCDFs-induced skin lesions. Chemosphere 2006;63:1410–8.

Uhm YK, Yoon SH, Kang IJ, Chung JH, Yim SV, Lee MH. Association of glutathione S-transferase gene polymorphisms (GSTM1 and GSTT1) of vitiligo in Korean population. Life Sci. 2007;1:223–227.

Ulusoy G, Arinç E, Adali O. Genotype and allele frequencies of polymorphic CYP2E1 in the Turkish population. Arch Toxicol. 2007;81(10):711-8.

Uno T, Obe Y, Ogura C, Goto T, Yamamoto K, Nakamura M, Kanamaru K, Yamagata H, Imaishi H. Metabolism of 7-ethoxycoumarin, safrole, flavanone and hydroxyflavanone by cytochrome P450 2A6 variants. Biopharm Drug Dispos. 2013 Mar;34(2):87-97.

Upton A, Johnson N, Sandy J, Sim E. Arylamine N-acetyltransferases - of mice, men and microorganisms. Trends Pharmacol Sci. 2001;22(3):140-6.

Uslu A, Ogus C, Ozdemir T, Bilgen T, Tosun O, Keser I. The effect of CYP1A2 gene polymorphisms on Theophylline metabolism and chronic obstructive pulmonary disease in Turkish patients. BMB Rep. 2010;43(8):530-4.

Vallone PM, Butler JM. AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques. 2004;37:226-231.

van der Hel OL, Bueno-de-Mesquita HB, van Gils CH, Roest M, Slothouber B, Grobbee DE, Peeters PH. Cumulative genetic defects in carcinogen metabolism may increase breast cancer risk (The Netherlands). Cancer Causes Control. 2005 Aug;16(6):675-81.

van der Merwe N, Bouwens CS, Pienaar R, van der Merwe L, Yako YY, Geiger DH, Kotze MJ. CYP2D6 genotyping and use of antidepressants in breast cancer patients: test development for clinical application. Metab Brain Dis. 2012;27(3):319-26.

van Emburgh BO, Hu JJ, Levine EA, Mosley LJ, Perrier ND, Freimanis RI, Allen GO, Rubin P, Sherrill GB, Shaw CS, Carey LA, Sawyer LR, Miller MS.

Polymorphisms in CYP1B1, GSTM1, GSTT1 and GSTP1, and susceptibility to breast cancer. Oncol Rep. 2008 May;19(5):1311-21.

van Lieshout EM, Roelofs HM, Dekker S, Mulder CJ, Wobbes T, Jansen JB, Peters WH. Polymorphic expression of the glutathione S-transferase P1 gene and its susceptibility to Barrett's esophagus and esophageal carcinoma. Cancer Res. 1999 Feb 1;59(3):586-9.

Vangsted AJ, Søeby K, Klausen TW, Abildgaard N, Andersen NF, Gimsing P, Gregersen H, Vogel U, Werge T, Rasmussen HB. No influence of the polymorphisms CYP2C19 and CYP2D6 on the efficacy of cyclophosphamide, thalidomide, and bortezomib in patients with Multiple Myeloma. BMC Cancer. 2010;10:404.

Vansina J (1984). Paths in the rainforest. Towards a history of political tradition in equatorial Africa. Currey: London.

Varela-Lema L, Ruano-Ravina A, Juiz Crespo MA, Kelsey KT, Loidi L, Barros-Dios JM. CYP1A1, mEH, and GSTM1 Polymophisms and Risk of Oral and Pharyngeal Cancer: A Spanish Case-Control Study. J Oncol. 2008;2008:741310.

Vural B, Yakar F, Derin D, Saip P, Yakar A, Demirkan A, Karabulut A, Ugurel E, Cine N, Kilicaslan Z, Tüzün E, Ozbek U. Evaluation of glutathione S-transferase P1 polymorphisms (Ile105Val and Ala114Val) in patients with small cell lung cancer. Genet Test Mol Biomarkers. 2012;16(7):701-6.

Wang J, Deng Y, Cheng J, Ding J, Tokudome S. GST genetic polymorphisms and lung adenocarcinoma susceptibility in a Chinese population. Cancer Lett. 2003;201(2):185-93.

Wang J, Jiang J, Zhao Y, Gajalakshmi V, Kuriki K, Suzuki S, et al. Genetic polymorphisms of glutathione S-transferase genes and susceptibility to colorectal cancer: a case-control study in an Indian population. Cancer Epidemiol. 2011;35(1):66-72.

Wang H, Zhang Z, Han S, Lu Y, Feng F, Yuan J. CYP1A2 rs762551 polymorphism contributes to cancer susceptibility: a meta-analysis from 19 case-control studies. BMC Cancer. 2012 Nov 19;12:528.

Watanabe J, Hayashi S, Kawajiri K. Different regulation and expression of the human CYP2E1 gene due to the RsaI polymorphism in the 5'-flanking region. J Biochem. 1994;116(2):321-6.

Watanabe J, Hayashi S, Nakachi K, Imai K, Suda Y, Sekine T, et al. PstI and RsaI RFLPs in complete linkage disequilibrium at the CYP2E gene. Nucleic Acids Res. 1990;18(23):7194.

Watson MA, Stewart RK, Smith GB, Massey TE, Bell DA. Human glutathione S-transferase P1 polymorphisms: relationship to lung tissue enzyme activity and population frequency distribution. Carcinogenesis. 1998;19(2):275-80.

Wentink MQ, Räkers M, Stupart DA, Algar U, Ramesar R, Goldberg PA. Incidence and histological features of colorectal cancer in the Northern Cape Province, South Africa. S Afr J Surg. 2010;48(4):109-13.

Wickliffe JK, Abdel-Rahman SZ, Lee C, Kormos-Hallberg C, Sood G, Rondelli CM, Grady JJ, Desnick RJ, Anderson KE. CYP1A2*1F and GSTM1 alleles are associated with susceptibility to porphyria cutanea tarda. Mol Med. 2011;17(3-4):241-7.

Wild CP, Yin F, Turner PC, Chemin I, Chapot B, Mendy M, et al. Environmental and genetic determinants of aflatoxin-albumin adducts in the Gambia. Int J Cancer. 2000;86(1):1-7.

Williamson K, Blench R (2000). Niger-Congo. In: Heine B, Nurse D, editors. African languages. Cambridge: Cambridge University Press. p11–42.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet. 2005;13(7):867-76.

Wooding SP, Watkins WS, Bamshad MJ, Dunn DM, Weiss RB, Jorde LB. DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. Am J Hum Genet. 2002 Sep;71(3):528-42.

Wu MT, Huang SL, Ho CK, Yeh YF, Christiani DC. Cytochrome P450 1A1 MspI polymorphism and urinary 1-hydroxypyrene concentrations in coke-oven workers. Cancer Epidemiol Biomarkers Prev. 1998 Sep;7(9):823-9.

Xia L, Gao J, Liu Y, Wu K. Significant association between CYP1A1 T3801C polymorphism and cervical neoplasia risk: a systematic review and meta-analysis. Tumour Biol. 2013 Feb;34(1):223-30.

Xu C, Rao YS, Xu B, Hoffmann E, Jones J, Sellers EM, et al. An in vivo pilot study characterizing the new CYP2A6*7, *8, and *10 alleles. Biochem Biophys Res Commun. 2002;290(1):318-24.

Yalin S, Hatungil R, Tamer L, Ates NA, Dogruer N, Yildirim H, et al. Glutathione S-transferase gene polymorphisms in Turkish patients with diabetes mellitus. Cell Biochem Funct. 2007;25:509–513.

Yamaguti GG, Lourenco GJ, Costa FF, Lima CS. High risk of ‘de novo’ acute myeloid leukaemia in individuals with cytochrome P450 A1 (CYP1A1) and NAD(P)H:quinone oxidoreductase 1 (NQO1) gene defects. Eur J Haematol. 2009;83:270–272.

Yang J, Qian LX, Wu HF, Xu ZQ, Sui YG, Wang XR, et al. Genetic polymorphisms in the cytochrome P450 1A1 and 2E1 genes, smoking, drinking and prostate cancer susceptibility: a case-control study in a Han nationality population in Southern China. Int J Urol. 2006;13(6):773-80.

Yen-Revollo JL, Van Booven DJ, Peters EJ, Hoskins JM, Engen RM, Kannall HD, et al. Influence of ethnicity on pharmacogenetic variation in the Ghanaian population. Pharmacogenomics J. 2009;9(6):373-9.

Yoshida R, Nakajima M, Watanabe Y, Kwon JT, Yokoi T. Genetic polymorphisms in human CYP2A6 gene causing impaired nicotine metabolism. Br J Clin Pharmacol. 2002, 54(5):511-7.

Yoshida K, Osawa K, Kasahara M, Miyaishi A, Nakanishi K, Hayamizu S, Osawa Y, Tsutou A, Tabuchi Y, Shimada E, Tanaka K, Yamamoto M, Takahashi J. Association of CYP1A1, CYP1A2, GSTM1 and NAT2 gene polymorphisms with colorectal cancer and smoking. Asian Pac J Cancer Prev. 2007;8(3):438-44.

Yuliwulandari R, Sachrowardi Q, Nishida N, Takasu M, Batubara L, Susmiarsih TP, Rochani JT, Wikaningrum R, Miyashita R, Miyagawa T, Sofro AS, Tokunaga K. Polymorphisms of promoter and coding regions of the arylamine N-acetyltransferase 2 (NAT2) gene in the Indonesian population: proposal for a new nomenclature. J. Hum. Genet. 2008;53:201-209.

Zaahl MG, Winter TA, Warnich L, Kotze MJ. The -237C-->T promoter polymorphism of the SLC11A1 gene is associated with a protective effect in relation to inflammatory bowel disease in the South African population. Int J Colorectal Dis. 2005;1-7.

Zgheib NK, Mitri Z, Geryess E, Noutsi P. Cytochrome P4502E1 (CYP2E1) genetic polymorphisms in a Lebanese population: frequency distribution and association with morbid diseases. Genet Test Mol Biomarkers. 2010;14(3):393-7.

Zihlif M, Imraish A, Irshaid YM. Frequency of certain single-nucleotide polymorphisms and duplication of CYP2D6 in the Jordanian population. Genet Test Mol Biomarkers. 2012 Oct;16(10):1201-5.

APPENDICES

Appendix A: Ethics clearance certificates

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Professor Himla Soodyall

CLEARANCE CERTIFICATE	**M090576**
PROJECT	Human Genetic Diversity and Disease (Previously M980553)
INVESTIGATORS	Professor Himla Soodyall.
DEPARTMENT	Diversity and Disease Research Unit
DATE CONSIDERED	09.05.29
DECISION OF THE COMMITTEE*	Annual Renewal Approved

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable
cc: Supervisor : Professor T Jenkins

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...

Appendix A continued

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Makkan

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER M060325**
PROJECT	Genetic Polymorphisms at Loci Implicated in Cancer Susceptibility among Sub-Saharan African Populations
INVESTIGATORS	Mr H Makkan
DEPARTMENT	Department of Human Genetics
DATE CONSIDERED	06.03.31
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 06.03.04 **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof H Soodyall

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix A continued

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Makkan

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER M060325**
PROJECT	Genetic Polymorphisms at Loci Implicated in Cancer Susceptibility among Sub-Saharan African Populations
INVESTIGATORS	Mr H Makkan
DEPARTMENT	Department of Human Genetics
DATE CONSIDERED	06.03.31
DECISION OF THE COMMITTEE*	Extension Approved

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 03/04/2013 **CHAIRPERSON** .....
(Professor P E Cleaton Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof H Soodyall

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix B: Comparative data

Table A) Comparative data - CYP1A1*2A and CYP1A1*2C allele frequencies

Region & country of sample	Population name	CYP1A1*2A rs4646903T>C		CYP1A1*2C rs1048943A>G		Reference
		Wild type	Mutant	Wild type	Mutant	
Southern Africa						
South Africa	South African Bantu-speakers	—	—	192 (100%)	0 (0%)	Masimirembwa et al., 1998
Eastern Africa						
Zimbabwe	Zimbabwean	—	—	296 (100%)	0 (0%)	Masimirembwa et al., 1998
Western Africa						
Benin	Benin	—	—	94 (100%)	0 (0%)	Jiang et al., 2005
Ghana	Ghanaian	290 (88.8%)	37 (11.2%)	300 (91.9%)	27 (8.1%)	Gronau et al., 2003
Ivory Coast	Ivorian	—	—	82 (61.7%)	51 (38.3%)	Santovito et al., 2010
Mali	Malian	88 (76%)	28 (24%)	116 (100%)	0 (0%)	Garte et al., 1998
Nigeria	Nigerian	335 (76.5%)	103 (23.5%)	—	—	Okobia et al., 2005
Northern Africa						
Egypt	Egyptian	—	—	381 (73.3%)	139 (26.7%)	El-Shennawy et al., 2011
Lybian	Tuaregs	101 (95.3%)	5 (4.7%)	104 (98.1%)	2 (1.9%)	Martinez-Labarga et al., 2007
Tunisia	Tunisian	248 (89.7%)	29 (10.3%)	—	—	Souiden et al., 2012
Europe						
Czech Republic	Slavic	47 (92.7%)	4 (7.3%)	54 (96.6%)	2 (3.4%)	Gemignani et al., 2007
Czech Republic	Slavic	587 (91%)	58 (9%)	—	—	Šarmanová et al., 2001
Greece	Greek	188 (75.5%)	61 (24.5%)	—	—	Agorastos et al., 2007
Hungary	Hungarian	95 (92.9%)	7 (7.1%)	110 (96.4%)	4 (3.6%)	Gemignani et al., 2007
Italy	Italian	—	—	742 (95.6%)	35 (4.4%)	Canova et al., 2010
Italy	Italian	—	—	447 (95.8%)	20 (4.2%)	D'Alo et al., 2004
Poland	Polish	168 (92.9%)	13 (7.1%)	196 (96.5%)	7 (3.5%)	Gemignani et al., 2007
Poland	Polish	610 (94.7%)	34 (5.3%)	633 (98.3%)	11 (1.7%)	Seremak-Mrozikiewicz et al., 2005
Romania	Romanian	76 (92.8%)	6 (7.2%)	89 (96.5%)	3 (3.5%)	Gemignani et al., 2007
Russia	Russian	60 (92.9%)	5 (7.1%)	69 (96.4%)	3 (3.6%)	Gemignani et al., 2007
Russia	Russian	937 (85.2%)	163 (14.8%)	—	—	Khvostova et al., 2012
Slovakia	Slovakians	50 (91.7%)	5 (8.3%)	58 (96.4%)	2 (3.6%)	Gemignani et al., 2007
Spain	Spanish	126 (89.6%)	15 (10.4%)	125 (88.9%)	16 (11.1%)	Esteller et al., 1997
Spain	Spanish	590 (89%)	73 (11%)	658 (96%)	28 (4%)	Landi et al., 2005
Spain	Spanish	1399 (88.6%)	180 (11.4%)	—	—	Lopez-Cima et al., 2012
Turkey	Turkish	252 (85.4%)	43 (14.6%)	—	—	Aydin-Sayitoglu et al., 2006
Asia						
Bangladesh	Bangladeshi	156 (70%)	67 (30%)	176 (79.3%)	46 (20.7%)	Islam et al., 2012
China	Chinese	252 (60%)	168 (40%)	344 (77.3%)	101 (22.7%)	Chen et al., 2007
China	Chinese	304 (60.8%)	196 (39.2%)	388 (77.6%)	112 (22.4%)	Yang et al., 2006
India	Indian	32 (64%)	18 (36%)	37 (74%)	13 (26%)	Kiruthiga et al., 2011
India	Indian	403 (80.6%)	97 (19.4%)	—	—	Parveen et al., 2010
Japan	Japanese	510 (60.6%)	332 (39.4%)	624 (74.1%)	218 (25.9%)	Kiyohara et al., 2012
Japan	Japanese	145 (60.2%)	96 (39.8%)	—	—	Nishino et al., 2008
Japan	Japanese	126 (69%)	57 (31%)	153 (81%)	36 (19%)	Sugawara et al., 2003
Japan	Japanese	114 (61%)	73 (39%)	145 (77.5%)	42 (22.5%)	Yoshida et al., 2007
Korea	Korean	116 (55.8%)	92 (44.2%)	—	—	Kim et al., 2004
Korea	Korean	—	—	1598 (75.5%)	518 (24.5%)	Kim et al., 2012
Korea	Korean	225 (62.2%)	137 (37.8%)	—	—	Kim et al., 2000
Thailand	Thai	551 (51.7%)	514 (48.3%)	732 (71.5%)	292 (28.5%)	Sangrajrang et al., 2009
Middle East						
Iran	Iranian	—	—	336 (88.5%)	44 (11.5%)	Razmkhah et al., 2011
Israel	Jewish	144 (87.5%)	21 (12.5%)	137 (83.2%)	28 (16.8%)	Guy et al., 2009
Lebanon	Lebanonese	159 (93.2%)	12 (6.8%)	—	—	Darazy et al., 2011
North America						
Canada	Canadian Caucasian	1584 (87.6%)	225 (12.4%)	—	—	Anderson et al., 2012
Canada	Canadian Caucasian	434 (92.4%)	36 (7.6%)	451 (96.1%)	19 (3.9%)	Krajinovic et al., 1999
United States of America	African-American	86 (98.9%)	1 (1.1%)	—	—	Olshan et al., 2000
United States of America	African-American	369 (76%)	117 (24%)	477 (96.2%)	19 (3.8%)	Taioli et al., 1998
United States of America	African-American	410 (76%)	129 (24%)	523 (97%)	16 (3%)	Garte et al., 1998
United States of America	Caucasian American	262 (96%)	11 (4%)	—	—	Olshan et al., 2000
United States of America	Caucasian American	116 (82.5%)	25 (17.5%)	130 (92%)	11 (8%)	Wickliffe et al., 2011
Mexico	Mexican	—	—	180 (66%)	92 (34%)	Gallegos-Arreola et al., 2004
Mexico	Mexican	—	—	174 (46%)	208 (54%)	Martínez-Ramírez et al., 2012
Mexico	Mexican	—	—	241 (46%)	288 (54%)	Perez-Morales et al., 2008
Costa Rica	Costa Rican	1238 (68.6%)	567 (31.4%)	—	—	Cornelis et al., 2004
South America						
Brazil	Brazilian	234 (91.1%)	23 (8.9%)	216 (84%)	41 (16%)	Gaspar et al., 2002
Chile	Chilean	160 (63%)	94 (37%)	122 (67.8%)	58 (32.2%)	Roco et al., 2012

Table B) Comparative data - CYP1A2*1C and CYP1A2*1F allele frequencies

Region & country of sample	Population name	CYP1A2*1C rs2069514G>A		CYP1A2*1F rs762551C>A		Reference
		Wild type	Mutant	Wild type	Mutant	
Southern Africa						
South Africa	South African Bantu-speakers	—	—	600 (61%)	383 (39%)	Dandara et al., 2011
Namibia	Ovambo	—	—	96 (54%)	82 (46%)	Fujihara et al., 2007
Eastern Africa						
Zimbabwe	Zimbabwean	—	—	124 (43.4%)	162 (56.6%)	Dandara et al., 2004
Tanzania	Tanzanian	—	—	73 (51.4%)	69 (48.6%)	Dandara et al., 2004
Ethiopia	Ethiopian	—	—	87 (50.4%)	86 (49.6%)	Aklillu et al., 2004
Western Africa						
Benin	Benin	—	—	40 (43%)	54 (57%)	Jiang et al., 2005
Nothern Africa						
Egypt	Egyptian	198 (93.4%)	14 (6.6%)	68 (31.8%)	145 (68.2%)	Hamdy et al., 2003
Tunisia	Tunisian	—	—	101 (50.5%)	99 (49.5%)	B'chir et al., 2009
Europe						
Czech Republic	Slavic	53 (98.8%)	1 (1.2%)	39 (68.2%)	18 (31.8%)	Gemignani et al., 2007
Denmark	Danish	—	—	28 (24.8%)	85 (75.2%)	Christiansen et al., 2000
Hungary	Hungarian	109 (99%)	1 (1%)	78 (67.2%)	38 (32.8%)	Gemignani et al., 2007
Italy	Italian	—	—	264 (34.6%)	498 (65.4%)	Canova et al., 2010
Netherlands	Dutch	—	—	73 (72.5%)	28 (27.5%)	Ketelslegers et al., 2006
Norway	Norwegian	—	—	114 (28.5%)	286 (71.5%)	Saebo et al., 2008
Poland	Polish	192 (98.9%)	2 (1.1%)	139 (67.5%)	67 (32.5%)	Gemignani et al., 2007
Romania	Romanian	87 (98.9%)	1 (1.1%)	63 (67.8%)	30 (32.2%)	Gemignani et al., 2007
Russia	Russian	68 (99%)	1 (1%)	49 (67.2%)	24 (32.8%)	Gemignani et al., 2007
Russia	Russian	—	—	524 (30.9%)	1174 (69.1%)	Khvostova et al., 2012
Slovakia	Slovakians	57 (99%)	1 (1%)	41 (67.1%)	20 (32.9%)	Gemignani et al., 2007
Spain	Spanish	612 (98.2%)	12 (1.8%)	229 (33.5%)	454 (66.5%)	Landi et al., 2005
Switzerland	Swiss	196 (98%)	4 (2%)	64 (32%)	136 (68%)	Todesco et al., 2003
Turkey	Turkish	—	—	124 (47.1%)	139 (52.9%)	Altayli et al., 2009
Turkey	Turkish	193 (96.2%)	8 (3.8%)	130 (49.2%)	134 (50.8%)	Uslu et al., 2010
Asia						
China	Chinese	—	—	160 (37.8%)	263 (62.2%)	Chen et al., 2007
China	Chinese	—	—	79 (32.6%)	162 (67.4%)	Han et al., 2001
India	Indian	373 (93.3%)	27 (6.8%)	226 (56.5%)	174 (43.5%)	Singh et al., 2010
Japan	Japanese	—	—	298 (37%)	508 (63%)	Shimada et al., 2009
Japan	Japanese	133 (74.3%)	46 (25.7%)	65 (37.1%)	110 (62.9%)	Yoshida et al., 2007
Korea	Korean	—	—	171 (68.4%)	79 (31.6%)	Fujihara et al., 2007
Mongolia	Mongolian	—	—	121 (78.8%)	33 (21.2%)	Fujihara et al., 2007
Thailand	Thai	—	—	236 (22.8%)	799 (77.2%)	Sangrajrang et al., 2009
North America						
Canada	Canadian Caucasian	—	—	558 (30.7%)	1260 (69.3%)	Anderson et al., 2012
Canada	Canadian Caucasian	—	—	117 (28.3%)	295 (71.7%)	Kotsopoulos et al., 2007
United States of America	African-American	—	—	338 (58.6%)	239 (41.4%)	Rebbeck et al., 2007
United States of America	Caucasian American	—	—	633 (54.4%)	531 (45.6%)	Rebbeck et al., 2007
United States of America	Caucasian American	—	—	45 (31.6%)	96 (68.4%)	Wickliffe et al., 2011
Costa Rica	Costa Rican	—	—	574 (31.8%)	1232 (68.2%)	Cornelis et al., 2004
South America						
Brazil	Brazilian	—	—	265 (34%)	513 (66%)	Shimada et al., 2009
Chile	Chilean	—	—	59 (23.1%)	195 (76.9%)	Roco et al., 2012

Table C) Comparative data - CYP2A6*7 and CYP2A6*8 allele frequencies

Region & country of sample	Population name	CYP2A6*7 rs5031016A>G		CYP2A6*8 rs28399468C>A		Reference
		Wild type	Mutant	Wild type	Mutant	
Western Africa						
Ghana	Ghanaian	420 (100%)	0 (0%)	420 (100%)	0 (0%)	Gyamfi et al., 2005
Europe						
Sweden	Swedish	380 (100%)	0 (0%)	380 (100%)	0 (0%)	Djordjevic et al., 2012
Asia						
China	Chinese	320 (93%)	24 (7%)	339 (98.6%)	5 (1.5%)	Nurfadhlina et al., 2006
India	Indian	348 (100%)	0 (0%)	345 (99.1%)	3 (0.9%)	Nurfadhlina et al., 2006
Japan	Japanese	118 (93.7%)	8 (6.3%)	124 (98.4%)	2 (1.6%)	Xu et al., 2002
Japan	Japanese	172 (93.5%)	12 (6.5%)	180 (97.8%)	4 (2.2%)	Yoshida et al., 2002
Korea	Korean	256 (88.9%)	32 (11.1%)	286 (99.3%)	2 (0.7%)	Djordjevic et al., 2012
Korea	Korean	403 (96.4%)	15 (3.6%)	412 (98.6%)	6 (1.4%)	Yoshida et al., 2002
Malaysia	Malaysian	517 (95.7%)	23 (4.3%)	513 (95%)	27 (5%)	Nurfadhlina et al., 2006
Thailand	Thai	228 (95%)	12 (5%)	240 (100%)	0 (0%)	Peamkrasatam et al., 2006
North America						
Canada	Canadian African-American	226 (100%)	0 (0%)	226 (100%)	0 (0%)	Mwenifumbo et al., 2005
Canada	Canadian Caucasian	220 (100%)	0 (0%)	220 (100%)	0 (0%)	Mwenifumbo et al., 2005
Canada	Canadian Caucasian	602 (100%)	0 (0%)	602 (100%)	0 (0%)	Xu et al., 2002

D) Comparative data - CYP2D6*3 and CYP2D6*4 allele frequencies

Region & country of sample	Population name	CYP2D6*3 rs35742686delA		CYP2D6*4 rs3892097G>A		Reference
		Wild type	Mutant	Wild type	Mutant	
Southern Africa						
South Africa	South African Bantu-speakers	152 (100%)	0 (0%)	147 (96.7%)	5 (3.3%)	Dandara et al., 2001
South Africa	South African Coloured	—	—	120 (96.8%)	4 (3.2%)	vd Merwe et al., 2012
South Africa	South African White	—	—	38 (76%)	12 (24%)	vd Merwe et al., 2012
Eastern Africa						
Zimbabwe	Zimbabwean	228 (100%)	0 (0%)	223 (98%)	5 (2%)	Dandara et al., 2001
Tanzania	Tanzanian	387 (99.7%)	1 (0.3%)	381 (98.2%)	7 (1.8%)	Dandara et al., 2001
Tanzania	Tanzanian	—	—	129 (95.9%)	6 (4.1%)	Hodel et al., 2012
Malawi	Malawian	—	—	25 (98%)	1 (2%)	Brown et al., 2012
Western Africa						
Ghana	Ghanaian	519 (100%)	0 (0%)	485 (93.4%)	34 (6.6%)	Griese et al., 1999
Ghana	Ghanaian	—	—	250 (76.5%)	77 (23.5%)	Gronau et al., 2003
Northern Africa						
Tunisia	Tunisian	—	—	172 (90.7%)	17.5 (9.3%)	Ouerhani et al., 2008
Europe						
Czech Republic	Slavic	441 (98.9%)	5 (1.1%)	344 (77.1%)	102 (22.9%)	Buzkova et al., 2008
Denmark	Danish	343 (98.7%)	5 (1.3%)	272 (78.1%)	76 (21.9%)	Vangsted et al., 2010
Germany	German	—	—	1178 (100%)	0 (0%)	Sachse et al., 1997
Greece	Greek	553 (97.7%)	13 (2.3%)	466 (82.3%)	100 (17.7%)	Arvanitidis et al., 2007
Poland	Polish	142 (97.9%)	3 (2.1%)	112 (76.9%)	34 (23.1%)	Gawronska-Szklarz et al., 1999
Portugal	Portuguese	99 (98.6%)	1 (1.4%)	87 (86.7%)	13 (13.3%)	Correia et al., 2009
Spain	Spanish	—	—	240 (87.1%)	36 (12.9%)	Gonzalez et al., 1998
Spain	Spanish	208 (99.1%)	2 (1%)	181 (86.2%)	29 (13.8%)	Menoyo et al., 2006
Switzerland	Swiss	114 (100%)	0 (0%)	82 (71.9%)	32 (28.1%)	Hersberger et al., 2000
Turkey	Turkish	—	—	175 (66.3%)	89 (33.7%)	Altayli et al., 2009
Turkey	Turkish	290 (98.3%)	5 (1.7%)	250 (84.7%)	45 (15.3%)	Aydin-Sayitoglu et al., 2006
United Kingdom	English	931 (98.1%)	18 (1.9%)	909 (79.6%)	233 (20.4%)	Smith et al., 1992
Asia						
Cambodia	Cambodian			74 (100%)	0 (0%)	Hodel et al., 2012
China	Chinese	446 (100%)	0 (0%)	445 (99.8%)	1 (0.2%)	Ji et al., 2002
India	Indian	447 (100%)	0 (0%)	414 (92.7%)	33 (7.3%)	Naveen et al., 2006
India	Indian	—	—	402 (80.4%)	98 (19.6%)	Parveen et al., 2010
Japan	Japanese	296 (100%)	0 (0%)	292 (98.6%)	4 (1.4%)	Furuno et al., 2001
Middle East						
Iran	Iranian	99 (99%)	1 (1%)	91 (91%)	9 (9%)	Hashemi-Soteh et al., 2011
Iran	Iranian	—	—	175 (87.5%)	25 (12.5%)	Kouhi et al., 2009
Israel	Jewish	—	—	136 (82.6%)	29 (17.4%)	Guy et al., 2009
Jordan	Jordanian	—	—	168 (87.2%)	25 (12.8%)	Zihlif et al., 2012
Saudi Arabia	Saudi Arabian	—	—	195 (96.5%)	7 (3.5%)	McLellan et al., 1997
North America						
Canada	Canadian Caucasian	457 (97.3%)	13 (2.7%)	391 (81.7%)	88 (18.3%)	Krajinovic et al., 1999
United States of America	African-American	489 (99.4%)	3 (0.6%)	456 (92.7%)	36 (7.3%)	Leathart et al., 1998
United States of America	American Indian	373 (99.7%)	1 (0.3%)	296 (79.1%)	78 (20.9%)	Fohner et al., 2013
United States of America	Caucasian American	917 (98.8%)	11 (1.2%)	760 (81.9%)	168 (18.1%)	Leathart et al., 1998
United States of America	Caucasian American	372 (98.4%)	6 (1.6%)	277 (73.3%)	101 (26.7%)	Skrętkowicz et al., 2011
Mexico	Mexican	829 (99.6%)	3 (0.4%)	794 (95.5%)	38 (4.5%)	Salazar-Flores et al., 2012
Panama	Panamanian	105 (100%)	0 (0%)	87 (82.9%)	18 (17.1%)	Jorge et al., 1999
South America						
Argentina	Argentinian	234 (97.5%)	6 (2.5%)	199 (82.9%)	41 (17.1%)	Lavandera et al., 2005
Brazil	Brazilian	373 (86.2%)	60 (13.8%)	415 (96.1%)	17 (3.9%)	Silviera et al., 2010
Chile	Chilean	251 (99%)	3 (1%)	223 (88.1%)	30 (11.9%)	Roco et al., 2012
Colombia	Colombian	136 (100%)	0 (0%)	117 (86%)	19 (14%)	Jorge et al., 1999

E) Comparative data - CYP2E1*5B (*Pst*I) and CYP2E1*5B (*Rsa*I) allele frequencies

Region & country of sample	Population name	CYP2E1*5B (PstI) rs3813867G>C		CYP2E1*5B (RsaI) rs2031920G>A		Reference
		Wild type	Mutant	Wild type	Mutant	
Southern Africa						
South Africa	South African Bantu-speakers	331 (100%)	0 (0%)	322 (97.3%)	9 (2.7%)	Chelule et al., 2006
Northern Africa						
Egypt	Egyptian	—	—	231 (98.3%)	4 (1.7%)	Hamdy et al., 2002
Europe						
Czech Republic	Slavic	43 (97.8%)	1 (2.2%)	54 (97.5%)	1 (2.5%)	Gemignani et al., 2007
Czech Republic	Slavic	—	—	605 (98%)	13 (2%)	Šarmanová et al., 2001
Hungary	Hungarian	89 (98%)	2 (2%)	109 (97.5%)	3 (2.5%)	Gemignani et al., 2007
Poland	Polish	156 (98%)	3 (2%)	194 (97.5%)	5 (2.5%)	Gemignani et al., 2007
Romania	Romanian	71 (97.9%)	2 (2.1%)	89 (97.5%)	2 (2.5%)	Gemignani et al., 2007
Russia	Russian	56 (98%)	1 (2%)	68 (97.5%)	2 (2.5%)	Gemignani et al., 2007
Slovakia	Slovakians	47 (98%)	1 (2%)	57 (97.6%)	1 (2.4%)	Gemignani et al., 2007
Spain	Spanish	261 (94.7%)	15 (5.3%)	—	—	Gonzalez et al., 1998
Spain	Spanish	622 (97.2%)	18 (2.8%)	568 (97.7%)	14 (2.3%)	Landi et al., 2005
Spain	Spanish	413 (95.9%)	18 (4.1%)	—	—	Rodrigo et al., 1999
Turkey	Turkish	—	—	281 (95.3%)	14 (4.7%)	Aydin-Sayitoglu et al., 2006
Turkey	Turkish	412 (100%)	0 (0%)	404 (98.1%)	8 (1.9%)	Ulusoy et al., 2007
Asia						
China	Chinese	—	—	358 (82.2%)	78 (17.8%)	Chen et al., 2007
China	Chinese	489 (81.6%)	110 (18.4%)	465 (77.4%)	136 (22.6%)	Huo et al., 2012
China	Chinese	385 (77%)	115 (23%)	385 (77%)	115 (23%)	Yang et al., 2006
India	Indian	283 (97.8%)	7 (2.2%)	283 (97.8%)	7 (2.2%)	Balaji et al., 2011
India	Indian	100 (100%)	0 (0%)	—	—	Mittal et al., 2005
Japan	Japanese	185 (76.8%)	56 (23.2%)	—	—	Nishino et al., 2008
Korea	Korean	119 (78.3%)	33 (21.7%)	—	—	Kim et al., 2004
Korea	Korean	294 (81.1%)	69 (18.9%)	—	—	Kim et al., 2000
Taiwan	Taiwanese	—	—	538 (78.6%)	147 (21.4%)	Hildesheim et al., 1997
Taiwan	Taiwanese	241 (75.2%)	80 (24.8%)	—	—	Hsieh et al., 2007
Thailand	Thai	—	—	865 (84.9%)	154 (15.1%)	Sangrajrang et al., 2010
Middle East						
Lebanon	Lebanonese	216 (100%)	0 (0%)	213 (98.6%)	3 (1.4%)	Zgheib et al., 2010
North America						
Canada	Canadian Caucasian	1745 (96.4%)	65 (3.6%)	—	—	Anderson et al., 2012
United States of America	Caucasian American	115 (84.6%)	21 (15.4%)	—	—	Wickliffe et al., 2011
Mexico	Mexican	—	—	497 (73.5%)	179 (26.5%)	Gordillo-Bastidas et al., 2010
South America						
Chile	Chilean	155 (85.8%)	26 (14.2%)	155 (85.8%)	26 (14.2%)	Roco et al., 2012

F) Comparative data - GSTM1*0 and GSTT1*0 genotype frequencies

Region & country of sample	Population ¥	Presence of both GSTM1*0 & GSTT1*0	Presence of GSTM1*0	Presence of GSTT1*0	Total Sample	References ₭
Southern Africa						
South Africa & Nambia	SEB [1] €		49 (21.8%)	66 (33.4%)		1; 2
South Africa	SAC [1]	9 (13%)	14 (20%)	39 (57%)	69	2
Namibia	OVA [1]	4 (3%)	15 (11.2%)	48 (35.8%)	134	3
Eastern Africa						
Zimbabwe	Zimbabwean €		36 (24.3%)	32 (26%)		1
Ethiopia	Ethiopia		67 (43.8%)	57 (37.3%)	153	4
Somalia	Somalia		40 (40%)	44 (44%)	100	5
Tanzania	Tanzania €		65 (31.4%)	35 (25%)		1
Central Africa						
Cameroon	Cameroonian		35 (2.8%)	59 (4.7%)	1265	4
Western Africa						
Gambia	Gambian €		68 (20.2%)	121 (37.1%)		6
Ivory Coast	Ivorian	19 (14.3%)	48 (36.1%)	44 (33.1%)	133	7
Northern Africa						
Egypt	Egyptian		206 (55.5%)	109 (29.4%)	371	8; 9
Tunisia	Tunisian	39 (21%)	118 (63.4%)	69 (37.1%)	186	10
Europe						
Austria	Austrian	19 (11.1%)	98 (57.3%)	30 (17.5%)	171	11
Czech Republic	Czech €		445 (52.2%)	141 (16%)		12
Denmark	Danish		105 (52.5%)	28 (14%)	200	5
France	French		37 (66.1%)	14 (25%)	56	13
Germany	German		1846 (52.5%)	621 (17.7%)	3513	14
Greece	Greek		148 (59.9%)	107 (43.3%)	247	15
Italy	Italian		59 (49.2%)	34 (28.3%)	120	4
Netherlands	Dutch		733 (53.1%)	499 (36.2%)	1380	16
Turkey	Turkish		269 (54.8%)	103 (21%)	491	27; 28
Spain	Spainish		1637 (53.4%)	645 (21%)	3066	17; 4; 18
Asia						
China	Chinese		952 (54.6%)	755 (43.3%)	1743	19; 20
India	Indian		300 (30%)	178 (17.8%)	999	21; 22
Japan	Japanese €		1110 (50.1%)	753 (49.6%)		23
Korea	Korean €		1952 (52.7%)	1853 (50.9%)		23
Mongolia	Mongolian	25 (12.1%)	96 (46.4%)	53 (25.6%)	207	3
Taiwan	Taiwanese		322 (56.1%)	287 (50%)	574	24
Middle East						
Bahrain	Bahraini	24 (14.4%)	83 (49.7%)	48 (28.7%)	167	10
Iran	Iranian		142 (37.4%)	73 (19.2%)	380	25
Lebanon	Lebanonese	23 (16.3%)	74 (52.5%)	53 (37.6%)	141	10
Saudi Arabia	Saudi Arabian	158 (22.7%)	371 (53.4%)	271 (39%)	695	26
North America						
Greenland	Greenlander		47 (47%)	46 (46%)	100	5
America	Caucasian American		991 (52.5%)	309 (16.4%)	1886	29; 30; 31
America	African-American		215 (28.6%)	145 (19.3%)	753	29; 30; 31
Canadia	Canadian Caucasian		269 (56.5%)	75 (15.8%)	476	32
Mexico	Mexican		143 (37.4%)	58 (15.2%)	382	33
South America						
Brazil	Brazilian		137 (31.5%)	128 (29.4%)	436	34; 35
Paraguay	Paraguayian		24 (35.8%)	12 (17.9%)	67	36

¥ For population name and ethnicity refer to table 1.

[1] symbolises a population from other studies. These populations are similar to those collected for this study therefore the same popultation code has been used

€ Null mutations were either typed individually or samples typed differed for each mutation, therefore the total sample size has not been displayed.

₭ Reference numbers - references can be found in accompanying reference table

Table F continued - references to GSTM1*0 and GSTT1*0 comparative data

Reference number	Reference
1	Dandara et al., 2002
2	Adams et al., 2003
3	Fujihara et al., 2009
4	Piacentini et al., 2011
5	Buchard et al., 2007
6	Wild et al., 2000
7	Santovito et al., 2010
8	Amer et al., 2011
9	Abdel-Rahman et al., 2012
10	Abdel-Halim et al., 2011
11	Gundacker et al., 2007
12	Sarmanova et al., 2001
13	Iarmarcovai et al., 2006
14	Rudolph et al., 2012
15	Agorastos et al., 2007
16	Van Der Hel et al., 2005
17	Cantor et al., 2010
18	Lopez-Cima et al., 2012
19	Wang et al., 2003
20	Liu et al., 2009
21	Kiran et al., 2008
22	Wang et al., 2011
23	Kurose et al., 2012
24	Tsai et al., 2006
25	Rafiee et al., 2010
26	Al-Dayel et al., 2008
27	Aydin-Sayitoglu et al., 2006
28	Yalin et al., 2007
29	Olshan et al., 2000
30	Millikan et al., 2000
31	Van Emburgh et al., 2008
32	Fischer et al., 2012
33	Martínez-Ramírez et al., 2012
34	Gatta´s et al., 2004
35	Magno et al., 2009
36	Gaspar et al., 2002

G) Comparative data - GSTP1-Ile105Val and GSTP1-Ala114Val allele frequencies

Region & country of sample	Population name	GSTP1*Ile105Val rs1695A>G		GSTP1*Ala114Val rs1138272G>A		Reference
		Wild type	Mutant	Wild type	Mutant	
Southern Africa						
South Africa	South African Bantu-speakers	48 (47%)	54 (53%)	—	—	Adams et al., 2003
South Africa	South African Bantu-speakers	203 (62.1%)	124 (37.9%)	283 (86.5%)	44 (13.5%)	Li et al., 2010
South Africa	South African Bantu-speakers	77 (87.9%)	11 (12.1%)	—	—	Masimirembwa et al., 1998
South Africa	South African Bantu-speakers	378 (47.2%)	422 (52.8%)	—	—	Matejcic et al., 2012
South Africa	South African Coloured	116 (60.5%)	76 (39.5%)	170 (89%)	21 (11%)	Li et al., 2010
South Africa	South African Coloured	366 (55.6%)	292 (44.4%)	—	—	Matejcic et al., 2012
Eastern Africa						
Zimbabwe	Zimbabwean	74 (75.8%)	24 (24.2%)	—	—	Masimirembwa et al., 1998
Tanzania	Tanzanian	86 (84.3%)	16 (15.7%)	—	—	Dandara et al., 2002
Western Africa						
Ghana	Ghanaian	420 (50.2%)	417 (49.8%)	—	—	Yen-Revollo et al., 2009
Nigeria	Yoruba	56 (62.2%)	34 (37.8%)	—	—	International HapMap 3 Consortium, 2010
Northern Africa						
Tunisia	Tunisian	513 (61.7%)	319 (38.3%)	352 (100%)	0 (0%)	Lakhdar et al., 2010
Europe						
Czech Republic	Slavic	36 (69.2%)	16 (30.8%)	53 (91.6%)	5 (8.4%)	Gemignani et al., 2007
Czech Republic	Slavic	456 (67.8%)	217 (32.2%)	—	—	Šarmanová et al., 2001
Denmark	Danish	563 (65.9%)	292 (34.1%)	864 (90.3%)	93 (9.7%)	Schwartzbaum et al., 2007
Finland	Finnish	199 (67.3%)	97 (32.7%)	305 (90.3%)	33 (9.7%)	Schwartzbaum et al., 2007
Hungary	Hungarian	74 (68.9%)	33 (31.1%)	106 (91%)	11 (9%)	Gemignani et al., 2007
Italy	Italian	178 (71%)	73 (29%)	237 (94.6%)	14 (5.4%)	Ballerini et al., 2003
Italy	Italian	527 (69.4%)	233 (30.6%)	—	—	Canova et al., 2010
Netherlands	Dutch	643 (63.5%)	370 (36.5%)	918 (91.4%)	87 (8.6%)	Dura et al., 2013
Netherlands	Dutch	66 (65.5%)	35 (34.5%)	95 (94.5%)	6 (5.5%)	Ketelslegers et al., 2006
Poland	Polish	130 (68.9%)	59 (31.1%)	189 (91.2%)	18 (8.8%)	Gemignani et al., 2007
Romania	Romanian	59 (69%)	27 (31%)	86 (91.3%)	8 (8.7%)	Gemignani et al., 2007
Russia	Russian	46 (68.9%)	21 (31.1%)	67 (91%)	7 (9%)	Gemignani et al., 2007
Slovakia	Slovakians	39 (68.8%)	18 (31.2%)	56 (90.9%)	6 (9.1%)	Gemignani et al., 2007
Spain	Spanish	434 (69.4%)	191 (30.6%)	650 (94.8%)	36 (5.2%)	Landi et al., 2005
Spain	Spanish	1035 (65.6%)	542 (34.4%)	—	—	Lopez-Cima et al., 2012
Sweden	Swedish	582 (65.9%)	302 (34.1%)	893 (90.3%)	96 (9.7%)	Schwartzbaum et al., 2007
Turkey	Turkish	189 (71.9%)	74 (28.1%)	—	—	Altayli et al., 2009
Turkey	Turkish	299 (75.9%)	95 (24.1%)	361 (91.6%)	33 (8.4%)	Vural et al., 2012
United Kingdom	English	588 (65.9%)	304 (34.1%)	902 (90.3%)	97 (9.7%)	Schwartzbaum et al., 2007
Asia						
China	Chinese	363 (82%)	80 (18%)	—	—	Chen et al., 2007
China	Chinese	190 (82.3%)	41 (17.7%)	—	—	Wang et al., 2003
India	Indian	187 (69.8%)	81 (30.2%)	—	—	Mittal et al., 2005
India	Indian	393 (78.6%)	107 (21.4%)	—	—	Parveen et al., 2010
India	Indian	628 (70.1%)	268 (29.9%)	837 (93.5%)	59 (6.5%)	Wang et al., 2011
Japan	Japanese	714 (84.9%)	127 (15.1%)	—	—	Kiyohara et al., 2012
Taiwan	Taiwanese	1077 (82.2%)	233 (17.8%)	—	—	Su et al., 2011
Thailand	Thai	777 (73.5%)	280 (26.5%)	—	—	Sangrajrang et al., 2010
Middle East						
Iran	Iranian	390 (68.2%)	182 (31.8%)	—	—	Hashemi et al., 2012
North America						
United States of America	African-American	51 (57.4%)	38 (42.6%)	—	—	Olshan et al., 2000
United States of America	African-American	80 (58%)	58 (42%)	109 (97.3%)	3 (2.7%)	Watson et al., 1998
United States of America	Caucasian American	179 (64.6%)	98 (35.4%)	—	—	Olshan et al., 2000
United States of America	Caucasian American	193 (67.1%)	95 (32.9%)	103 (91.2%)	10 (8.8%)	Watson et al., 1998
Mexico	Mexican	198 (51.8%)	184 (48.2%)	—	—	Martínez-Ramírez et al., 2012
South America						
Argentina	Argentinian	66 (64.7%)	36 (35.3%)	—	—	Galvan et al., 2011
Brazil	Brazilian	269 (63.2%)	157 (36.8%)	—	—	Coutino et al., 2010
Brazil	Brazilian	112 (73.7%)	40 (26.3%)	—	—	Da Silva et al., 2008

H) Comparative data - NAT2*14A allele frequencies

Region & country of sample	Population name	NAT2*14A rs1801279G>A		Reference
		Wild type	Mutant	
Southern Africa				
South Africa	South African Bantu-speakers	234 (97.5%)	6 (2.5%)	Adams et al., 2003
South Africa	South African Bantu-speakers	171 (89.1%)	21 (10.9%)	Dandara et al., 2003
South Africa	South African Coloured	162 (95.3%)	8 (4.7%)	Adams et al., 2003
Eastern Africa				
Zimbabwe	Zimbabwean	281 (86.2%)	45 (13.8%)	Dandara et al., 2003
Tanzania	Tanzanian	207 (88.5%)	27 (11.5%)	Dandara et al., 2003
Tanzania	Tanzanian	121 (85.5%)	21 (14.5%)	Hodel et al., 2012
Kenya	Luhya	163 (91.6%)	15 (8.4%)	International HapMap 3 Consortium, 2010
Kenya	Maasai	275 (96.2%)	11 (3.8%)	International HapMap 3 Consortium, 2010
Ethiopia	Ethiopian	54 (100%)	0 (0%)	Luca et al., 2008
Somalia	Somali	100 (100%)	0 (0%)	Bayoumi et al., 1997
Somalia	Somali	47 (97.9%)	1 (2.1%)	Upton et al., 2001
Sudan	Sudanese	264 (97.1%)	8 (2.9%)	Bayoumi et al., 1997
Central Africa				
Cameroon	Baka Pygmies	58 (96.7%)	2 (3.3%)	Patin et al., 2006a
Cameroon	Bakola Pygmies	76 (95%)	4 (5%)	Patin et al., 2006a
Cameroon	Bedzan Pygmies	61 (95.3%)	3 (4.7%)	Patin et al., 2006b
Cameroon	Ngumba Bantus	27 (84.4%)	5 (15.6%)	Patin et al., 2006b
Central African Republic	Biaka Pygmies	43 (89.6%)	5 (10.4%)	Patin et al., 2006b
Democratic Republic of Congo	Mbuti Pygmies	37 (77.1%)	11 (22.9%)	Patin et al., 2006b
Gabon	Gabonese	95 (91.3%)	9 (8.7%)	Deloménie et al., 1996
Western Africa				
Ghana	Ghanaian	718 (89.8%)	82 (10.3%)	Yen-Revollo et al., 2009
Mali	Dogons	123 (94.6%)	7 (5.4%)	Deloménie et al., 1996
Nigeria	Yoruba	82 (91.1%)	8 (8.9%)	International HapMap 3 Consortium, 2010
Senegal	Senegalese	95 (90%)	11 (10%)	Toure et al., 2012
Northern Africa				
Morocco	Moroccan	88 (100%)	0 (0%)	Upton et al., 2001
Europe				
Czech Republic	Slavic	2173 (99.9%)	3 (0.1%)	Cascorbi et al., 1999
Finland	Finnish	96 (100%)	0 (0%)	Patin et al., 2006a
France	French	120 (100%)	0 (0%)	Deloménie et al., 1996
Italy	Italian	167 (99.4%)	1 (0.6%)	International HapMap 3 Consortium, 2010
Poland	Polish	496 (100%)	0 (0%)	Mrozikiewicz et al., 1996
Russia	Russian	580 (100%)	0 (0%)	Gaikovitch et al., 2003
Sardinia	Sardinian	98 (100%)	0 (0%)	Upton et al., 2001
Siberia	Siberian	72 (100%)	0 (0%)	Fuselli et al., 2007
Spain	Spanish	679 (99.9%)	1 (0.1%)	Landi et al., 2005
Sweden	Swedish	100 (100%)	0 (0%)	Upton et al., 2001
Turkey	Turkish	606 (100%)	0 (0%)	Aynacioglu et al., 1997
United Kingdom	English	224 (100%)	0 (0%)	Hein et al., 2002
Asia				
Cambodia	Cambodian	118 (97.5%)	3 (2.5%)	Hodel et al., 2012
China	Chinese	424 (100%)	0 (0%)	Song et al., 2009
India	Indian	122 (100%)	0 (0%)	Lin et al., 1994
Indonesia	Indonesian	424 (100%)	0 (0%)	Yuliwulandari et al., 2007
Japan	Japanese	158 (100%)	0 (0%)	Lin et al., 1994
Korea	Korean	576 (100%)	0 (0%)	Lee et al., 2003
Taiwan	Taiwanese	200 (100%)	0 (0%)	Lin et al., 1994
Thailand	Thai	88 (100%)	0 (0%)	Patin et al., 2006a
Vietnam	Vietnamese	144 (100%)	0 (0%)	Cavaco et al., 2007
Republic of the Philippines	Filipinos	200 (100%)	0 (0%)	Lin et al., 1994
Middle East				
Iran	Iranian	458 (100%)	0 (0%)	Torkaman-Boutorabi 2007
Oman	Omani	352 (100%)	0 (0%)	Al-Moundhri et al., 2007
Saudi Arabia	Saudi Arabian	974 (100%)	0 (0%)	Bu et al., 2004
North America				
Canada	Canadian Caucasian	1815 (99.8%)	3 (0.2%)	Anderson et al., 2012
Canada	Canadian Indian	962 (100%)	0 (0%)	Hegele et al., 2000
United States of America	African-American	330 (92.1%)	28 (7.9%)	Lin et al., 1994
United States of America	Caucasian American	196 (100%)	0 (0%)	Lin et al., 1994
Mexico	Mexican	32 (100%)	0 (0%)	Fuselli et al., 2007
Nicaragua	Nicaraguans	273 (99.6%)	1 (0.4%)	Martinez et al., 1998
South America				
Brazil	Brazilian	775 (95.9%)	33 (4.1%)	Teixeira et al., 2007
Colombia	Colombian	51 (98.1%)	1 (1.9%)	Fuselli et al., 2007
Ecuador	Ecuadorian	30 (100%)	0 (0%)	Fuselli et al., 2007
Peru	Peruvian	300 (100%)	0 (0%)	Fuselli et al., 2007

Appendix C: SNP Marker regions with DNA Sequence, PCR primer and SBE primer sequence

- PCR fragment sequences:

  - β-Globin

    5'- gaagagccaaggacaggtacGGCTGTCATCACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAG GAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAAC TGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCA Aggtgaacgtggatgaagtt -3'

    SBE – POLYGACT + tggtgcatctgactcctgag

    Presence of G allele (CONTROL BAND)

  - GSTP1 — codon 105 polymorphism (Ile105Val / A to G)

    5'-accccagggctctatgggaaGGACCAGCAGGAGGCAGCCCTGGTGGACATGGTGAATGACGGCGTGGAGGACCTCCGCTGCAA ATACGTCTCCCTCATCTACACCAACTATGTGAGCATCTGCACCAGGGTTGGGCACTGGGGGCTGAACAAAGAaaggggcttcttgt gccctca-3'

    SBE – POLYGACT + aggacctccgctgcaaatac

    Presence of G allele (mutant)

    Presence of A allele (wild-type)

- GSTP1 — codon 114 polymorphism (Ala114Val / C to T)

5'-gttgtggggagcaagcagaggAGAATCTGGGACTCTGGTGTCTGGCCTGGGGCAGACGGGGGTGTCTCAGGGGCTGGGAGGGA
TGAGAGTAGGATGATACATGGTGGTGTCTGGCAGGAGGCGGGCAAGGATGACTATGTGAAGGCACTGCCCGGGCAACTGA
AGCCTTTTGAGACCCTGCTGTCCCAGAACCAgggaggcaagaccttcattgtg-3'

SBE (3' end) – POLYGACT + cttcacatagtcatccttgccc

Presence of A allele (mutant)

Presence of G allele (wild-type)

- CYP1A1*2A (3'-flanking region) — T to C

5'-cagtgaagaggtgtagccgcTGCACTTAAGCAGTCTGTTTGAGGGACAAGACTCTATTTTTTGAGACAGGGTCCCCAGGTCATCC
AGGCTGGAGTGCACTGGTACCATTTTGTTTCACTGTAACCTCCACCTCCTGGGCCCACACGATTCTCCCACCTCAGCCTCTGAG
TAGTTGGGGCCGCAGGCACACGCCACCACAGCTTTTTTTTTTTTTTTTTTTTTTTTTGTAGAGATGGGGTTTCACCATGTTGCC
CAGGCTGGTCTCAAACTCGCTCAAGTGATCCACCTGCCTCAGCCTCCCAAAGTGCTGGGATTACAggcatgagacaagactccta-3'

SBE – POLYGACT + tgtttcactgtaacctccacctcc

Presence of C/ A allele (mutant)

Presence of T allele (wild-type

- CYP1A1*2C in exon 7 — A to G

5'- cagctgtctccctctggttaCAGGAAGCTATGGGTCAACCCATCTGAGTTCCTACCTGAACGGTTTCTCACCCCTGATGGTGCTAT CGACAAGGTGTTAAGTGAGAAGGTGATTATCTTTGGCATGGGCAAGCGGAAGTGTATCGGTGAGACCATTGCCCGCTGGga ggtctttctcttcctggc -3

SBE – POLYGACT + ggaagtgtatcggtgagacc

Presence of G allele (mutant)

Presence of A allele (wild-type)

- CYP1A2*1C site in the 5' flanking region — G to A

5'- gctacacatgatcgagctatacATGACATATGCACTTTTCCATTTATTTATTTATTTTTGAGACAGAATCTTGCTCTGTCACCCAGG CTGGAGTGCAGTGGTGCGATCTTGGCTCACCGCAACCTCCGCCTCTCAgattcaagcaattgtcatgc-3'

SBE – POLYGACT + ccgcaacctccgcctctc

Presence of A allele (mutant)

Presence of G allele (wild-type)

- CYP1A2*1F site in intron 1 — C to A

5'- gatgatgtgtggaggagagaGCCAGCGTTCATGTTGGGAATCTTGAGGCTCCTTTCCAGCTCTCAGATTCTGTGATGCTCAAAG
GGTGAGCTCTGTGGGCACAGGACGCATGGTAGATGGAGCTTAGTCTTTCTGGTATCCAGCTGGGAGCCAAGCACAGAACAC
GCATCAGTGTTTATCAAATGACTGAGGAAATGAATGAATGaatgtctccatctcaaccct-3'

SBE – POLYGACT + aaagggtgagctctgtgggc

Presence of C allele (wild-type)

Presence of A allele (mutant)

- CYP2A6*7 (in green) and CYP2A6*8 (in purple) in 3' flanking region

5'- ctcaagtcctcccagtcaccTAAGGACACTGACGTGTCCCCCAAACACGTGGGCTTTGCCACGATCCCATAAACTACACCATG
AGCTTCCTGCCCCGCTGAGCGAGGGCTGTGCCGGTGCAGGTCTGGTGGGCGGGGCCAGGGAAAGGGCAGGGCCAAGACCG
GGCTTGGGAGAGGGGCGCAGCTAAGACTGGGGGCAGGATGGCGGAAAGGAAGGGGCGtggtggctagagggaagaga-3'

CYP2A6*7

SBE (3' end) – POLYGACT + acgtgtttgggggacacgtca

Presence of A allele (wild-type)

Presence of G allele (mutant)

CYP2A6*8

SBE (3' end) – POLYGACT + ggcaggaagctcatggtgtagttt

Presence of C allele (wild-type)

Presence of A allele (mutant)

- CYP2D6*3A (1bp del (A) in position 2549)

5'- taggtgctgaatgctgtccCCGTCCTCCTGCATATCCCAGCGCTGGCTGGCAAGGTCCTACGCTTCCAAAAGGCTTTCCTGACCCA GCTGGATGAGCTGCTAACTGAGCACAGGATGACCTGGGACCCAGCCCAGCCCCCCCGAGACCTGACTGAGGCCTTCCTGgcag agatggagaaggtgag -3'

SBE – POLYGACT + gatgagctgctaactgagcac

Presence of A allele (wild-type)

Presence of G allele (mutant)

- CYP2D6*4A (1846G>A)

5'- gtgggtgatgggcagaagGGCACAAAGCGGGAACTGGGAAGGCGGGGGACGGGGAAGGCGACCCCTTACCCGCATCTCCC ACCCCCAGGACGCCCCTTTCGCCCcaacggtctcttggacaaag-3'

SBE – POLYGACT + ccgcatctcccaccccca

Presence of G allele (wild-type)

Presence of A allele (mutant)

- CYP2E1*5B(*PstI*)

5'- cactggaaaggaaagagaggAGGAGGCGGCAGGCTAACCCACCCGTGAGCCAGTCGAGTCTACATTGTCAGTTCTCACCTCG AGGGGTGCCAAAAACCAGAGGGAAGCAAAGGCCCCTGAAGCCTCTGCCAGAGCCAACGCCCCTTCTTGGTTCAGGAGAGGT GCAGTGTTAGGTGCAGCACAACcaatgacttgcttatgtggc-3'

SBE – POLYGACT + cccttcttggttcaggagag

Presence of G allele (wild-type)

Presence of C allele (mutant)

- CYP2E1*5B(*Rsa*I)

5'- aagtgatttggctggattgtAAATGACTTTTATTTTCTTCATTTCTCATCATATTTTCTATTATACATAAAGATTCATTGTTAATA
TAAAAGTACAAAATTGCAACCTATGAATTAAGAACTTCTATATATTGCCAGTTAGAAGACAGAATGAAAAACATTCTCTTCAT
TCTAACCACACACACAAAAAAGCTCCACAAAATACCTATGGACTACCTtcatagaaggtggaagaggg– 3'

SBE (3' end) – POLYGACT + gttcttaattcataggttgcaatttt

Presence of G allele (wild-type)

Presence of A allele (mutant)

- NAT2*14A (G to A transition at position 191)

5'- ttaggggatcatggacattgAAGCATATTTTGAAAGAATTGGCTATAAGAACTCTAGGAACAAATTGGACTTGGAAACATTAAC
TGACATTCTTGAGCACCAGATCCGGGCTGTTCCCTTTGAGAACCTTAACATGCATTGTGGGCAAGCCATGGAGTTGGGCTTA
GAGGCTATTTTTGATCACATTGTAAGAAGAAACCGGGGTGG gtggtgtctccaggtcaatc-3'

SBE – POLYGACT + tgatcacattgtaagaagaaacc

Presence of G allele (wild-type)

Presence of A allele (mutant)

Appendix D: Pair-wise genetic distance (F_{st}) for indigenous populations

	CAR_U	CAR_P	DAMA	DRC	HER	HIM	JU	KHWE	MOZ	NAMA	OVA	SAC	SEB	UGA	XUN	ZAM
CAR_U	0.0000															
CAR_P	0.0288	0.0000														
DAMA	0.0294	0.1217	0.0000													
DRC	0.0063	0.0537	0.0464	0.0000												
HER	0.0134	0.0330	0.0276	0.0347	0.0000											
HIM	0.0056	0.0530	0.0124	0.0234	0.0056	0.0000										
JU	0.0630	0.0974	0.0878	0.0714	0.0313	0.0842	0.0000									
KHWE	0.0119	0.0346	0.0236	0.0310	0.0000	0.0114	0.0165	0.0000								
MOZ	0.0020	0.0499	0.0093	0.0195	0.0075	0.0000	0.0830	0.0109	0.0000							
NAMA	0.0450	0.0736	0.0752	0.0693	0.0139	0.0501	0.0272	0.0164	0.0489	0.0000						
OVA	0.0047	0.0426	0.0137	0.0259	0.0042	0.0000	0.0808	0.0087	0.0000	0.0431	0.0000					
SAC	0.0111	0.0123	0.0543	0.0288	0.0069	0.0209	0.0424	0.0074	0.0173	0.0156	0.0159	0.0000				
SEB	0.0109	0.0398	0.0204	0.0336	0.0000	0.0025	0.0369	0.0000	0.0034	0.0106	0.0009	0.0072	0.0000			
UGA	0.0124	0.0736	0.0140	0.0311	0.0261	0.0018	0.1121	0.0313	0.0000	0.0573	0.0019	0.0337	0.0172	0.0000		
XUN	0.0697	0.0980	0.0807	0.0799	0.0319	0.0774	0.0000	0.0182	0.0809	0.0502	0.0776	0.0555	0.0404	0.1177	0.0000	
ZAM	0.0058	0.0297	0.0254	0.0246	0.0080	0.0000	0.0796	0.0114	0.0000	0.0539	0.0000	0.0176	0.0078	0.0125	0.0742	0.0000