Detection of sequence diversity in the *CYP2C19* gene of Xhosa South African individuals: An analytical and comparative study including *in silico* and functional analysis of the 5' flanking region

BY

Britt Drögemöller

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science (MSc) in Genetics at Stellenbosch University

Supervisor: Prof Louise Warnich
Co-supervisor: Prof Dana Niehaus
Co-supervisor: Dr Renate Hillermann-Rebello

March 2010

DECLARATION:

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature:……..

Date:……..

SUMMARY

The prevalence of adverse drug reactions (ADR) and treatment failure in South Africa requires urgent addressing and it is the aim of pharmacogenetics to aid in the alleviation of these ADRs and treatment failures. However, considering the high level of genetic diversity present in African populations, preliminary analysis of the genetic profiles of South African populations is required before pharmacogenetics can be successfully implemented in the South African context. Therefore this study aimed to characterise the gene encoding the drug metabolising enzyme, CYP2C19, in the South African Xhosa population.

To identify the common *CYP2C19* sequence variation present in the Xhosa population, semi-automated sequence analysis of *CYP2C19* was performed on 15 healthy Xhosa individuals. The variation detected was then prioritised through various *in silico* analyses for further restriction fragment length polymorphism (RFLP) genotyping in an additional 85 healthy Xhosa individuals to confirm the frequencies of the prioritised variants in a larger cohort, while the copy number variation (CNV) present in the entire 100 Xhosa individuals was analysed with the use of duplex real-time PCR. To functionally validate the *in silico* data obtained for the 5'-upstream variants, dual luciferase reporter assays were utilised. In addition to these analyses, multi-species comparisons were used to highlight regions of high sequence similarity within the 5'-upstream regions, while CpG island analysis was utilised to identify possible CpG islands occurring within and around the *CYP2C* genes.

Sequence analysis of the *CYP2C19* gene revealed 30 variants, of which five were novel. Subsequent to RFLP analysis, the frequencies of the allele-defining variants detected in this population, namely *CYP2C19*2*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17* were found to be 0.21, 0.09, 0.09 and 0.10, respectively. Additionally, the novel non-synonymous V374I variant, which was designated *CYP2C19*28*, was found to occur at a frequency of 0.01. Dual luciferase reporter assays revealed that the construct containing the rs7902257 variant, demonstrated a significant decrease in the fold induction observed when compared to the "wild type" construct (P = 0.0077). This variant was designated *CYP2C19*27* and was detected at a frequency of 0.33 in the Xhosa population. In addition to this, multi-species comparisons revealed four highly conserved regions, all of which were present within LINE L1 repetitive elements. Although putative CpG islands were identified in and around the *CYP2C* genes, no direct correlations could be made between the differences in expression observed between the genes and the presence of the CpG islands. The role of these islands with regards to the epigenetic regulation of these genes therefore remains to be elucidated.

To our knowledge, this study provides the most comprehensive data for *CYP2C19* in a South African population and shows that the Xhosa population displays a unique genetic profile, which differs from those of other populations, including the Cape Mixed Ancestry population of South Africa. Thus, novel genotyping platforms need to be developed in order to successfully apply pharmacogenetics to the diverse populations residing in South Africa.

OPSOMMING

Die doel van Farmakogenetika is om daadwerklike aandag aan die hoë voorkoms van nadelige geneesmiddel reaksies en mislukte behandelings te skenk en om hierdie voorkoms in Suid-Afrika te verlaag. Die bevolkingsgroepe van Afrika het hoë vlakke van genetiese diversiteit en dus hang die susksesvolle toepassing van Farmakogenetika in Suid-Afrika af van die voorlopige analise van die genetiese profiele van die Suid-Afrikaanse bevolkingsgroepe. Om hierdie rede was die doel van hierdie studie om die geneesmiddel metaboliseerings geen, *CYP2C19*, in 'n Suid-Afrikaanse Xhosa bevolkingsgroep te karakteriseer.

Die *CYP2C19* volgorde van 15 Xhosa individue is bepaal om die algemene variasie teenwoordig in die *CYP2C19* geen te bevestig. Hierdie variasies is deur verskeie *in silico* analises geprioritiseer vir verder restriksie fragment lengte polimorfisme (RFLP) genotipering in 85 gesonde Xhosa individue om die frekwensie in 'n groter groep te bevestig, terwyl die kopie aantal variasie teenwoordig in hierdie 100 Xhosas geanaliseer is met Taqman® CNV toetse. Om die *in silico* data vir die 5'-stroomop variante funksioneel te bevestig, is daar gebruik gemaak van tweedeelige luciferase verklikker toetse. Verder is multi-spesie vergelykings gebruik om 5'-stroomop streke met hoë vlakke van ooreenstemming te identifiseer, terwyl CpG-eiland analise gebruik is om moontlike CpG-eilande in die omgewing van die *CYP2C* gene te identifiseer.

Met behulp van volgorde bepaling van die *CYP2C19* geen, is 30 variante geïdentifiseer. Uit hierdie variante was vyf vir die eerste keer met hierdie studie opgespoor. Met die gebruik van RFLP analise, is die alleel definerende variante naamlik *CYP2C19*2*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17*, teen 'n frekwensie van 0.21, 0.09, 0.09 en 0.10 in die Xhosa bevolkingsgroep gevind. Verder was die nie-sinonieme variant, V374I, wat vir die eerst keer geïdentifiseer en *CYP2C19*28* genoem is, teen 'n frekwensie van 0.01 gevind. Tweedeelige luciferase verklikker toetse het bewys dat die konstruk met die rs7902257 variant 'n beduidende afname in induksie in vergelyking met die "wilde tipe" konstruk gewys het (P = 0.0077). Hierdie variant was *CYP2C19*27* genoem en was teen 'n frekwensie van 0.33 in die Xhosa bevolking gevind. Die multi-spesie vergelykings het vier gekonserveerde streke geïdentifiseer wat in LINE L1 herhalende elemente gevind is. Alhoewel CpG-eilande in die omgewing van die *CYP2C* gene gevind is, kon geen direkte korrelasies gemaak word tussen die veranderinge in uitdrukking van die gene en die teenwoordigheid van die CpG-eilande nie. Die rol van hierdie eilande met betrekking tot epigenetiese regulasie van hierdie gene moet dus nog ontrafel word.

Tot ons kennis het hierdie projek die mees voledige inligting vir *CYP2C19* in ’n Suid-Afrikaanse bevolkingsgroep gegee en het bewys dat die Xhosa bevolkingsgroep ‘n unieke genetiese profiel vertoon, wat van ander bevolkingsgroepe, insluitend die Kaapse Gemenge Herkoms populasie van Suid-Afrika, verskil het. Indien farmokogenetika suksesvol in die diverse bevolkingsgroepe van Suid-Afrika toegepas kan word, moet daar gebruik gemaak word van nuwe genotipering metodes.

ACKNOWLEDGEMENTS

I would like to acknowledge and express my gratitude to the following people and institutions.

My supervisor Prof Warnich: for challenging, supporting and advising me, as well as providing me with the many tools and opportunities to learn and grow.

My co-supervisor Prof Niehaus: for his insight and for providing the opportunities to participate in the psychiatric side of this project.

My co-supervisor Dr Hillermann-Rebello: for her input in my thesis.

Galen Wright: for his endless support and innovative ideas. For challenging, inspiring, advising and encouraging me throughout.

Dr Mauritz Venter: for encouragement, as well as advise and help with all bioinformatic and promoter-based analysis.

Stefanie Malan: for DNA sequence analysis of exons 5-9 of *CYP2C19*.

Danielle Da Silva: for *CYP2C19* data on the Cape Mixed Ancestry population.

Marika Bosman: for assistance with the dual luciferase reporter assays.

Anthony La Grange: for statistical analysis performed in this study.

The rest of lab 231 (Jomien de Jager, Louise van der Merwe and Lundi Korkie): for all their assistance and for providing a happy work place.

Carel van Heerden, Rene Veikondis, Gloudi Agenbag and the rest of the Central Analytical Facility of Stellenbosch University: for advice and help with sequence analysis.

Prof Liezl Koen: for assistance with the psychiatric side of this project.

Dr Craig Kinnear: for assistance with the TaqMan® CNV Assays.

The Xhosa individuals: for their participation.

Harry Crossley Foundation and the National Research Foundation: for funding.

Joint Cold Spring Harbour/Wellcome Trust organising committee, SASHG organising committee and the Biological Psychiatry organising committee: for providing stipends to allow me to attend the various conferences.

My parents: for always believing in me, providing support and countless opportunities to allow me to reach the point that I have arrived at.

The rest of my friends and family: for playing an important role in my life.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: MATERIALS AND METHODS

CHAPTER 4: RESULTS

CHAPTER 5: DISCUSSION

LIST OF TABLES

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: MATERIALS AND METHODS

CHAPTER 4: RESULTS

LIST OF SYMBOLS AND ABBREVIATIONS

3’	3-prime end
5’	5-prime end
α	Alpha
+	And
ß	Beta
Δ	Change in
χ^2	Chi-squared
©	Copyright
°C	Degrees Celsius
$	Dollar
=	Equal to
γ	Gamma
>	Greater than
µg	Microgram
µl	Microlitre
µM	Micromolar
%	Percentage
±	Plus-minus
£	Pound
®	Registered trademark
<	Smaller than
3D	Three dimensional
X	Times
™	Trademark

A	Adenine
AAC	Associated ancestral clusters
ADRs	Adverse drug reactions
AIDS	Acquired Immunodeficiency Syndrome
Apo	Apolipoprotein
APS	Ammonium persulphate ($H_8N_2O_8S_2$)
ARMS	Amplification refractory mutation systems
ART	Antiretroviral therapy
ARV	Antiretroviral

AS-PCR	Allele-specific-polymerase chain reaction
ATCC	American Type Culture Collection
AUC	Area under the curve
BLAST	Basic local alignment search tool
bp	Base pair
BSA	Bovine serum albumin
c	Concentration
C	Cytosine
CAR	Constitutive androstane receptor
C/EBP	CCAAT/enhancer binding protein
CHOP	C/EBP homologous protein
CMA	Cape Mixed Ancestry
cm^2	Centimetres squared
CNV	Copy number variation
CO_2	Carbon dioxide
C_T	Cycle threshold
CYP	Cytochrome P450
df	Degrees of freedom
dH_2O	Distilled water
DME	Drug metabolising enzymes
DMEM	Dulbecco's Modified Eagle's Medium
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphates
Dr	Doctor
E	Exon
E. cloni	*Escherichia cloni*
EDTA	Ethylenediaminetetraacetic Acid ($C_{10}H_{16}N_2O_8$)
e.g.	*Exempli gratia*
EM	Extensive metaboliser
EMSA	Electrophoretic mobility shift assay
et al.	*Et al.ii*
EtBr	Ethidium bromide ($C_{10}H_{16}N_2O_8$)

etc.	*Et cetera*
F	Forward primer
FBS	Fetal bovine serum
FDA	Food and Drug Association
g	Gram
G	Guanine
gDNA	Genomic deoxyribonucleic acid
GRE	Glucocorticoid receptor
GST	Glutathione S-transferase
H	Histidine
H	Histone
HepG2	Human hepatocellular liver carcinoma cell line
HGP	Human Genome Project
HIV	Human Immunodeficiency Virus
HNF	Hepatic nuclear factor
H. pylori	*Helicobacter pylori*
hr	Hour
HWE	Hardy-Weinberg equilibrium
I	Isoleucine
ID	Identification
IM	Intermediate metaboliser
IVS	Intervening sequence
Kb	Kilobase
l	Litre
L	Leucine
LB	Luria-Bertani medium
LD	Linkage disequilibrium
LINE	Long interspersed repetitive element
LOD	Logarithm of odds
Ltd	Limited

Luc	Luciferase
m	Mutagenic primer
M	Molar
MAF	Minor allele frequency
max	Maximum
mg	Milligram
$MgCl_2$	Magnesium Chloride
Min	Minutes
miRNA	Micro ribonucleic acid
ml	Millilitre
mM	Millimolar
MR	Metabolic ratio
mRNA	Messenger ribonucleic acid
n	Sample size
NAT	N-acetyltransferase
NCBI	National Centre for Biotechnology Information
NF-kappaB	Nuclear factor kappa-B
Ng	Nanogram
NHS	National Health Service
nm	Nano metre
No	Number
NRF	National Research Foundation
nt	Nucleotide
Oct-1	Octamer binding protein-1
ORF	Open reading frame
P	Probability
ρ	Pico
PAA	Polyacrylamide
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
P+E1	Promoter region and exon one

PM	Poor metaboliser
PolyPhen	Polymorphism Phenotyping
PPI	Proton pump inhibitor
Prof	Professor
Pty	Proprietary limited company
PXR	Pregnane X receptor
q	Long arm of chromosome
R	Arginine
R	Rand
R	Reverse primer
REC	Research ethics committee
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
rpm	Revolutions per minute
rs	RefSNP
rVISTA	Rank Vista
s	Sequencing primer
SASHG	South African Society for Human Genetics
SDS	Sodium dodecyl sulfate ($C_{12}H_{25}OSO_3Na$)
Sec	Seconds
SEM	Standard error of the mean
SIFT	Sorting intolerant from tolerant
SINE	Short interspersed repetitive element
SMR	Standardized mortality rates
SNP	Single nucleotide polymorphism
svm	Support vector machine
T	Thymine
Taq	*Thermus aquaticus*
TBE	Tris borate ethylenediaminetetraacetic acid buffer
TD	Tardive dyskinesia
TE	Tris ethylenediaminetetraacetic acid buffer
TEMED	N,N,N',N'-tetramethylethylenediamine ($C_6H_{16}N_2$)

TFPGA	Tools for population genetic analysis
TPMT	Thiopurine S-methyltransferases
Tris	Tris(hydroxymethyl)aminomethane ($C_4H_{11}NO_3$)
U	Unit (enzyme quantity)
UGT	Uridine 5'-diphospho-glucuronosyltransferase
UK	United Kingdom
UM	Ultra rapid metaboliser
UNAIDS	The United Nations Joint Programme on HIV/AIDS
USA	United States of America
USF	Upstream stimulatory factor
UTR	Untranslated region
UV	Ultraviolet
v	Version
V	Volts
V	Valine
vs	Versus
v/v	Volume per volume
WHO	World Health Organization
w/v	Weight per volume
www	World wide web

CHAPTER 1:

INTRODUCTION

CHAPTER 1: INTRODUCTION

Since the beginning of time, the A-, T-, C- and Gs that constitute life, have been shuffled and re-shuffled to allow for the constant generation of a dynamic and colourful world. To crack this deceptively simple code has been, and continues to be, the goal of thousands of the geneticists all around the world. Aiding in this deciphering process, the 3.2 billion base pairs of DNA sequence obtained from the Human Genome Project (www.genome.gov/HGP), along with access to a wide variety of computational tools, allow for endless possibilities to unearth patterns, similarities and differences that may provide vital clues to the missing pieces in the puzzle of life. Where previously only coding regions were understood, a whole range of other exciting areas have begun to emerge that would have been virtually impossible to identify without the help of computers and high-throughput technology.

A further aspect of genetic studies that has provided a wealth of information is the investigation of genetic variation. In theory, the presence of a mere single nucleotide polymorphism (SNP) could lead to disastrous or advantageous consequences, depending on the context of the mutation. However, in practice, the role of genetic variation is not always quite as straightforward due to the fact that most diseases are complex and controlled by a number of different factors and genes (Davey Smith *et al.* 2005). Even so, the study of genetic variation is of immense importance to a current day understanding of living systems and each discovery can provide missing clues to bridge the ever narrowing gaps in the quest for a comprehensive understanding of biological systems.

Taking into consideration that at present not all six million of the validated SNPs found on the National Centre for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/snp/) can be tested in every individual; population genetics, haplotype analysis and various bioinformatic tools present methods by which to prioritise and sort through the vast amounts of data that are constantly generated. When focussing on the members of the species *Homo sapiens*, the history of their origin and migration throughout time, provides vital information with regards to their genetic make-up. Originally, approximately 200 000 years ago, the modern human was found in Africa, however groups of individuals began migrating out of the continent approximately 100 000 years ago. These individuals were subsequently separated and began residing in their respective areas, interbreeding with each other and forming populations (Campbell and Tishkoff 2008). As each of these populations shared, and in many cases continue to share, a gene pool, the alleles present in a specific population and their frequencies, are likely to differ from those found in another population (Klug and Cummings 2003). Thus, by studying representatives from particular populations, it is possible to determine which variants are likely to occur at significant frequencies in

those populations, thereby prioritising variants from the six million identified, that warrant studying in that particular population. To assist in population specific studies and determining how variants are inherited, the HapMap project has provided a readily available comparison of approximately 1.6 million SNPs from different populations by obtaining genotype information for 1 115 individual samples from 11 populations (Duan *et al.* 2008).

When examining the various populations of the world, African populations are of special interest as they are the most ancient of populations and their genetic make-up has not been widely studied. For this reason a substantial amount of valuable information can be obtained from these populations. According to the Out of Africa theory, the modern human originated in Africa (Tishkoff and Verrelli 2003) and it appears that Africans developed a population substructure within the continent before migrating to other parts of the world (Tishkoff *et al.* 1996; Garrigan *et al.* 2004; Harding and McVean 2004; Plagnol and Wall 2006; Garrigan *et al.* 2007; Yotova *et al.* 2007). These substructures were and remain, based on ethnic, linguistic, geographical and environmental factors. When a select group of Africans from a specific sub-group migrated to other parts of the world, a bottleneck effect was observed. This means that in these derivative populations today, a smaller number of variants are observed at a higher frequency. In contrast, African populations are older and larger and have been exposed to greater variation in climate, diet and exposure to infectious disease, thus showing greater diversity than non-African populations. It is therefore important to bear in mind, when examining African populations, that the size and age of these populations may result in high levels of within-population genetic variation (Reed and Tishkoff 2006; Campbell and Tishkoff 2008).

It is rather ironic that the ancient and diverse African populations which consist of more than 2 000 distinct ethno-linguistic groups (http://www.ethnologue.com) have not been widely studied with regards to genetic, and more specifically, pharmacogenetic research and most studies focussing on African populations have been exclusively performed on African-American individuals, which have originated predominantly from Western Africa (Tishkoff *et al*. 2009). Furthermore, it has recently been suggested that Southern African populations appear to show the greatest genetic diversity (Tishkoff *et al.* 2009), therefore highlighting a need to examine South African populations. When considering pharmacogenetic studies, currently, the Venda population is the only South African population for which the *CYP2C19* gene has been examined to our knowledge (Dandara *et al.* 2001; Matimba *et al.* 2009). As the Venda comprise only 2.3% of the South African population (http://www.statssa.gov.za/census01/HTML/default.asp), making them the second smallest population in the country, other studies on South African populations need to be performed. For this project we have chosen to focus on the Xhosa population, which comprise 17.6% (http://www.statssa.gov.za/census01/HTML/default.asp) of the South African population, making it

the second largest unique South African population. Therefore, this study which is aimed at examining the genetic diversity of the pharmacogenetically relevant *CYP2C19* gene in the Xhosa population will make an important contribution to our knowledge regarding the genetic profile of a Southern African population. This data may have valuable implications for the application of pharmacogenetics in Southern Africa.

CHAPTER 2:

LITERATURE REVIEW

CHAPTER 2: LITERATURE REVIEW

2.1 Pharmacogenetics

2.1.1 Background

"Pharmacogenetics" was termed by Vogel in 1959 to describe the inherited difference in response to therapeutic agents. Since then, specifically with the birth of molecular biology and the completion of the Human Genome Project, much research has been executed on the topic and many polymorphisms with pharmacogenetic relevance have been described (Manolopoulos 2007). These polymorphisms may occur within and around genes that code for drug metabolisers, receptors or transporters and were first described by Oscarson (2003) as monogenetic traits exhibiting more than one allele at the same locus, which exist stably in a population, producing more than one phenotype with regards to drug reaction.

For specific drug metabolising enzyme (DME) genes, individuals may exhibit one of four phenotypes with respect to drug metabolism, which are categorised according to enzyme functionality. The categories are poor metabolisers (PMs), with two non-functional copies of the gene; intermediate metabolisers (IMs), with two decreased function copies or one non-functional copy of the gene; extensive metabolisers (EMs), with two normal copies of the gene and ultra-rapid metabolisers (UMs) with gene duplications or increased function mutations, unaccompanied by non-functional mutations (McKinnon and Evans 2000, Dandara *et al.* 2001; Ingelman-Sundberg *et al.* 2007). Each of these classes metabolise drugs with varying efficiencies and therefore require different drug dosages. Figure 2.1 demonstrates the mechanism by which in the case of drugs that are inactivated by DMEs, the plasma concentration of an ingested drug continually increases in PMs, whereas it remains constant in EMs. The same principle can be applied to UMs, however in this case the plasma concentration of the drug will be lower rather than higher. Conversely, for drugs that are activated by DMEs, UM individuals will experience higher plasma concentrations of the activated drug. A low plasma concentration of the activated drug is generally associated with adverse drug reactions (ADRs) and the possible development of resistance to the drug as a result of sub-inhibitory concentrations, whereas a high plasma concentration may be associated with therapeutic failure (Gardiner and Begg 2006). Thus, pharmacogenetics can be divided up into safety and efficacy pharmacogenetics, which are aimed at decreasing ADRs and treatment failure, respectively (Roses 2004). Therefore the ultimate goal of pharmacogenetics is to ensure that the area under the curve (AUC) (refer to Figure 2.1) is equal for all individuals (Kirchheiner *et al.* 2005).

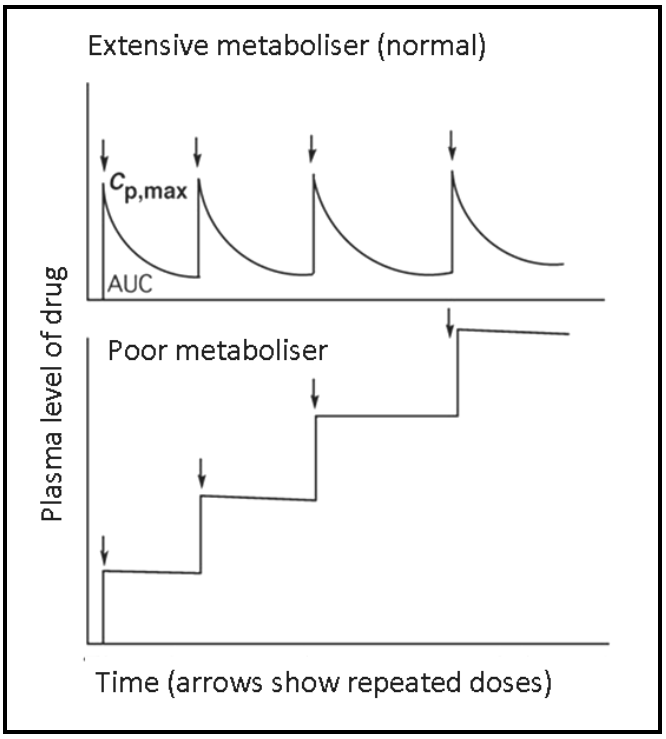


Figure 2.1: The difference in AUCs for PM and EM individuals with regards to the plasma level of drugs.
(Oritz de Montellano 2005) (Reprinted with permission from American Association of Pharmaceutical Scientists)

Phenotypically, drug metaboliser classes can be determined through the measurement of specific hydroxylation indices in the urine after the ingestion of a standard dose of probe drug relevant to the drug metabolising enzyme under inspection (Goldstein and De Morais 1994). By performing studies which correlate phenotypic and genotypic data, the reliability of pharmacogenetic data can be improved. After phenotypic validation, genotyping of the variants with pharmacogenetic application can be utilised to aid in the elimination of ADRs and the optimisation of drug dosage. Thus, as opposed to a trial and error based drug dosage prescription, a genotype test can be implemented in the treatment plan of the patient throughout his/her life.

Pharmacogenetics should be applied to drugs whose side effects or inefficiency significantly affect the well-being of the patients and the economy of the country and should primarily be applied to situations where treatment alleviation is essential. Furthermore, drugs with narrow therapeutic indices will reap the benefits of pharmacogenetics more obviously than general response drugs. It is, however, important to remember that although an ADR may not be severe, the comfort of a patient remains important and may influence the compliance and thus treatment outcome of that patient. The process by which pharmacogenetics should be studied and eventually implemented is depicted in Figure 2.2.

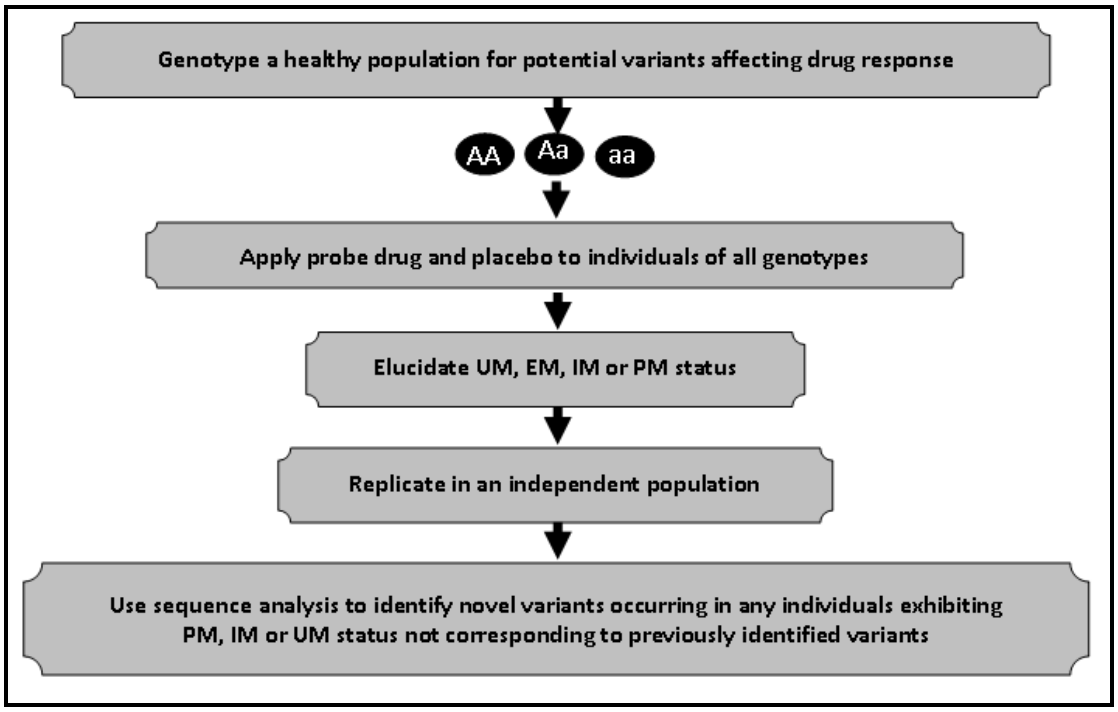


Figure 2.2: Implementation of pharmacogenetics.
(Adapted from Willard and Ginsburg 2009)

2.1.2 Combat of Adverse Drug Reactions (ADRs)

ADRs contribute significantly to economic burdens and health care quality throughout the world. In the USA more than two million cases of ADRs were reported to occur every year (Lazarou *et al.* 1998), while more recently it has been reported that the NHS in England require 1.6 million hospital bed days every year due to ADRs (Wiffen *et al.* 2002). Furthermore, it has been estimated that approximately £637 million is spent by the NHS on ADRs annually (Davies *et al.* 2009). In India and the United Kingdom, 6.85% and 6.5% of patients are hospitalised due to ADRs respectively, of which 59.62% and 72% are avoidable (Pirmohamed *et al.* 2004; Patel *et al.* 2007). Fatal ADRs were estimated to be the 7th leading cause of death in Sweden (Wester *et al.* 2008) and it has been estimated that most drugs are only effective in half of all patients (Allison 2008), which is of serious consequence considering that Americans have been reported to take on average 14.3 prescriptions a year (Cox *et al.* 2008). It is therefore clear that it would be highly advantageous for both economic and health reasons, to decrease the occurrence of ADRs and treatment failure, which are likely to be a frequent and severe consequence in third world countries, such as South Africa.

Among the ADRs reported to date, haemolysis (Carson *et al.* 1956), peripheral neuropathy (Hughes *et al.* 1954), severe skin rash (Calza *et al.* 2009) tardive dyskinesia (TD) (Arranz and de Leon 2007), cardiovascular effects and sudden death (Brown *et al.* 2004) have all been implicated as serious ADRs that require urgent addressing. On a more positive note, it has been estimated that through the

implementation of pharmacogenetics, the rate of ADRs could be reduced by 10-20% and the efficiency of drugs could be increased by 10-15% (Ingelman-Sundberg 2004). This can either be implemented through the exclusion of drugs which are metabolised by polymorphic enzymes, through individualised drug treatment based on genotype status (Ingelman-Sundberg *et al.* 1999) or through development of drugs which are metabolised by more than one enzyme (Ortiz de Montellano 2005).

It must, however, be acknowledged that the occurrence of ADRs cannot be attributed solely to genetic factors and that smoking, diet, concomitant drugs, physiological or disease status, age and demographic factors also play a large role (Sotanieui *et al.* 1997; Kashuba *et al.* 1998; Ingelman-Sundberg *et al.* 1999; Arranz and de Leon 2007). External factors often induce or saturate metabolic pathways which would otherwise work with greater efficiency. Often, these external factors influence the transcriptional activity of the drug metabolising enzymes, emphasizing the need to study and understand not only the coding regions of genes with pharmacogenetic application, but also their upstream regulatory regions (Dossing *et al.* 1983; Wilkins *et al.* 1987). Additionally concomitant factors have been shown to inhibit or induce therapeutic agents in a gene-dosage dependant manner, with UMs showing the most sensitivity, followed by EMs, IMs and lastly PMs (Caraco *et al.* 1995, 1996; Desta *et al.* 2002). Since different populations harbour different frequencies of UMs, EMs and PMs, it stands to reason that these different populations will show different sensitivities to external factors. Taking all of this into account, it is essential that the variants present in genes of pharmacogenetic value, as well as their surrounding areas, are extensively studied and understood. Data obtained from these studies should then be used in combination with the information available on the external factors influencing drug metabolism and patients should be carefully monitored by health care providers.

In order for the successful management of ADRs, an interdisciplinary approach in which technology, scientists, pharmaceutical companies, the government, health care providers and patients all work together to create a greater awareness, is required. All of these disciplines require suitable education on ADRs and access to the appropriate facilities in order to reduce the occurrence of ADRs. This begins in the clinical setting with both the patient and health care provider. In studies performed in Netherlands, Germany, Sweden and the United Kingdom, it has been reported that only 44-70% of ADRs are reported (Belton *et al.* 1995; Eland *et al.* 1999; Backstrom *et al.* 2000; Hasford *et al.* 2002), however this rate is even lower in African countries such as Nigeria, where only 16.4% of ADRs are reported (Okezie and Olufunmilayo 2008). Possible grounds for the poor documentation of ADRs in Nigeria are depicted in Figure 2.3. The lack of reliable information for the occurrence and rate of ADRs, combined with an ignorance regarding the serious consequences of

ADRs, further complicates the process of ADR elimination. By eliminating ADRs, health care costs can be reduced and patient compliance is expected to increase (Mabadeje *et al.* 1991).

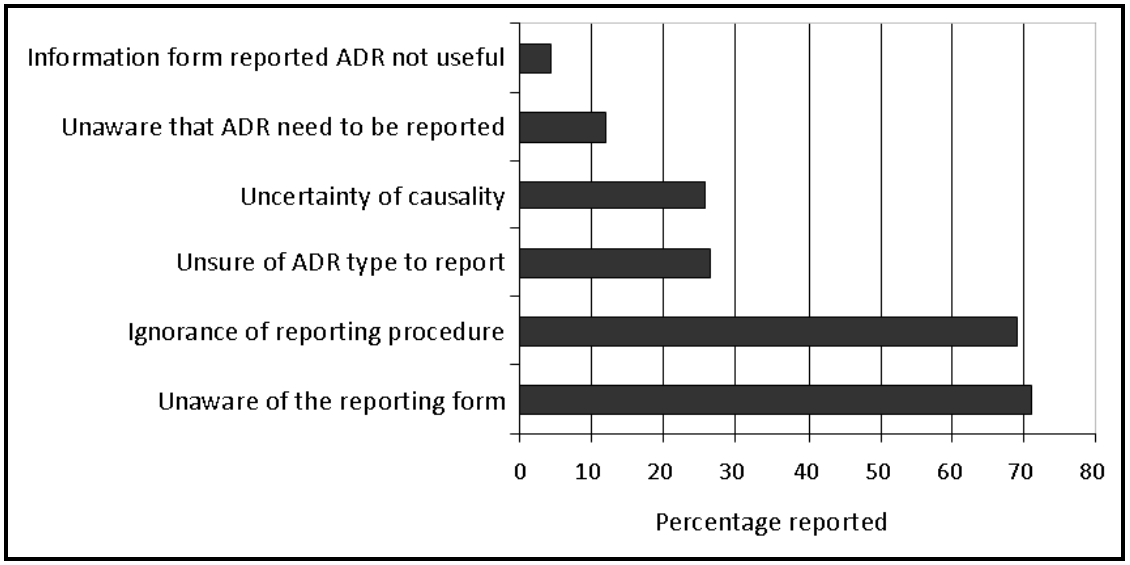


Figure 2.3: Factors that play a role in the under-reporting of ADRs by doctors in the Nigerian clinical setting. (Okezie and Olufunmilayo 2008) (Reprinted with permission from John Wiley and Sons)

2.2 The Drug Metabolising Enzymes

2.2.1 The Cytochrome (P450) Genes

As the human race has progressed, certain genes have evolved to better suit the environment and corresponding needs of the species. An excellent example of gene evolution is the cytochrome P450 (*CYP*) gene family. When evolutionary events are examined, it seems that as animals took to the land and began to eat plants, the *CYP* genes began to evolve more rapidly. This occurred due to the fact that *CYP* genes are responsible for metabolising toxins, including those derived from plants. As the animals consumed plants, the plants would evolve and create new toxins to defend themselves and in response, the *CYP* genes were forced to evolve. As a result, the human genome now contains 57 active *CYP* genes and 58 *CYP* pseudogenes (Ingelman-Sundberg *et al.* 2005). Furthermore, in certain populations where the *CYP* genes are less frequently required, the corresponding genes are less stringently protected from the accumulation of variants, often rendering non-functional genes. Similarly in populations where the genes are more frequently utilised, over-active genes are often observed. An interesting example of this is the high percentage (30%) of *CYP2D6* gene duplications observed in Ethiopians as opposed to the 5.3% of functional *CYP2D6* gene duplications observed in mixed European countries (Sistonen *et al.* 2009). It has been hypothesised that to prevent starvation, Ethiopians are required to ingest a larger variety of plant toxins, thus the need for CYP2D6 to metabolise these toxins is greater. In this case two copies of *CYP2D6* are more beneficial, whereas

two or even one copy of *CYP2D6* in more developed countries is often unnecessary (Ingelman-Sundberg 2005).

At first glance, the evolution of genes would not seem to be of any major consequence to humans. However, modern medicine has developed certain drugs with the assumption that the *CYP* genes in humans will remain functional and will consequently remove all toxins derived from the ingested drugs. As modern medicine has further developed, it has come to our attention that this assumption requires re-evaluation. The presence of non-functional genes may lead to toxic side-effects as a result of the ingested drug, further jeopardising the health of the patients, while UM genes may result in treatment failure. Thus, the screening of genes with pharmacogenetic applications for variants is of vital importance for the implementation of successful treatment plans.

For pharmacogenetics to be successfully implemented, a comprehensive understanding of the drug pathways must exist, including the absorption, distribution, metabolism and elimination of the drug. This study has placed its focus on the metabolism of the drugs. Drugs are predominantly metabolised in the liver by a process which is controlled by two phases. Phase I enzymes convert the drug into a metabolite, while Phase II enzymes inactivate the metabolite by coupling it to an endogenous substance. As far as clinically prescribed drugs are concerned, 80% of Phase I enzymes belong to the CYP family (Eichelbaum *et al.* 2006), whereas Phase II enzymes are represented by enzymes such as N-acetyltransferases (NATs), thiopurine S-methyltransferases, UDP glucuronosyltransferases (UGTs) and glutathione S-transferases (GSTs) (Arranz and de Leon 2007). Phase I enzymes are responsible, mainly through oxidation, for defending the body against endogenous agents such as steroids, fatty acids and prostaglandins as well as exogenous agents such as carcinogens, environmental pollutants and importantly in the context of pharmacogenetics, detoxifying drugs (Shimada *et al.* 1994; Prior *et al.* 1999). An inability to metabolise a drug efficiently may lead to a build up of the drug in the bloodstream which may in turn lead to serious toxic side effects as a result of drug ingestion (Prior *et al.* 1999; Dandara *et al.* 2001; Gaikovitch *et al.* 2003; Nakamoto *et al.* 2007). Alternatively an increased metabolism of the drug will lead to decreased drug affectivity. In cases where DMEs convert a prodrug into an active metabolite the opposite is true. By optimising drug dosage, extra costs and ADRs can be eliminated by the removal of unnecessarily high dosages of drugs, whereas therapeutic failure can be eliminated by the remedying of low drug dosages (Kirchheiner *et al.* 2001).

It has been estimated that the *CYP* genes are responsible for metabolising over 90% of currently prescribed drugs (Masimirembwa and Hasler 1997). Considering the vast number of *CYP* genes present in the human genome, a categorising system for these genes is essential. All CYP enzymes

sharing more than 40% amino acid sequence similarity, belong to the same family and are given the same Arabic numeral (e.g. CYP2); sub-families displaying more than 55% sequence similarity to each other are assigned common letters (e.g. CYP2C) and lastly individual enzymes are given an individual Arabic numeral (e.g. CYP2C19) (Levy 1995; Nelson *et al.* 1996).

Despite the large number of CYP genes that are present in the human genome, less than 10 appear to be important to pharmacogenetic applications (Oscarson 2003). Recently a committee, including the FDA, categorised enzymes according to their importance with regards to pharmacogenetic applications (http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126957). These categories included known "valid" pharmacogenomic biomarkers and "exploratory" pharmacogenomic biomarkers. The known "valid" pharmacogenomic biomarkers were those molecules expressing a measurable genetic polymorphism which was proven to be associated with a variable drug response. These molecules were CYP2D6, CYP2C19, CYP2C9, thiopurine S-methyltransferase (TPMT) and UGT1A1 (Andersson *et al.* 2005). With regards to the CYP family, when examining which of the genes have the most academic and industry related importance, Ingelman-Sundberg *et al.* (2007), determined which *CYP* gene websites were most frequently visited (refer to Figure 2.4). It appears that *CYP2D6*, *CYP2C9* and *CYP2C19* receive the most attention, in that order. This is, based among other things, on the gene variation present in these genes, which include gene duplications, gene deletions, amino acid changes and mutations (including those in non-coding regions) which result in non-functional enzyme products (Ingleman-Sundberg *et al.* 2007).

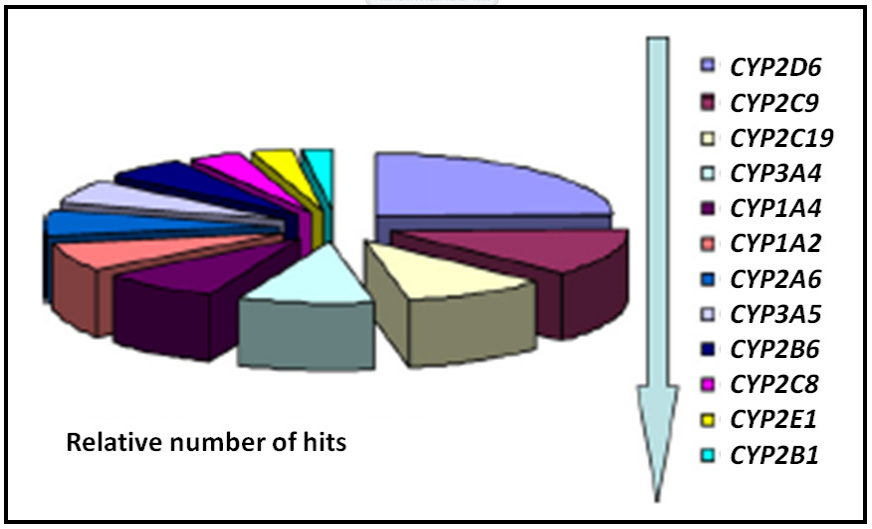


Figure 2.4: Popularity of the *CYP* genes as determined from the number of hits each gene receives on the *CYP* allele website.
(Ingelman-Sundberg *et al.* 2007) (Reprinted with permission from Elsevier Limited)

As technology develops and decreases in cost, the genotyping of *CYP* genes before the relevant drugs are prescribed becomes an ever increasing reality. With the arrival of microchip arrays where more

than a million SNPs can be genotyped easily for $1000 (Steemers and Gunderson 2007) the possibility of screening for a wide variety of gene variations seems to be a likely course of action in the clinical setting. A recent development with application for the CYP enzymes was the release of the first FDA approved pharmacogenetics test. This was the Roche AmpliChip P450 in 2004, with 27 *CYP2D6* alleles and three *CYP2C19* alleles (de Leon *et al.* 2006). Furthermore, Kirchheiner *et al.* (2005) have already provided dosage recommendations according to genotype (refer to Figure 2.5). It has been reported that of the 1 200 FDA approved drugs released between 1945-2005, 120 have pharmacogenomic information on their labels, of which 69 are human genomic biomarkers and 63% of these refer to the *CYP* genes (Frueh *et al.* 2008); illustrating the growth pharmacogenetics with special reference to the *CYP* genes.

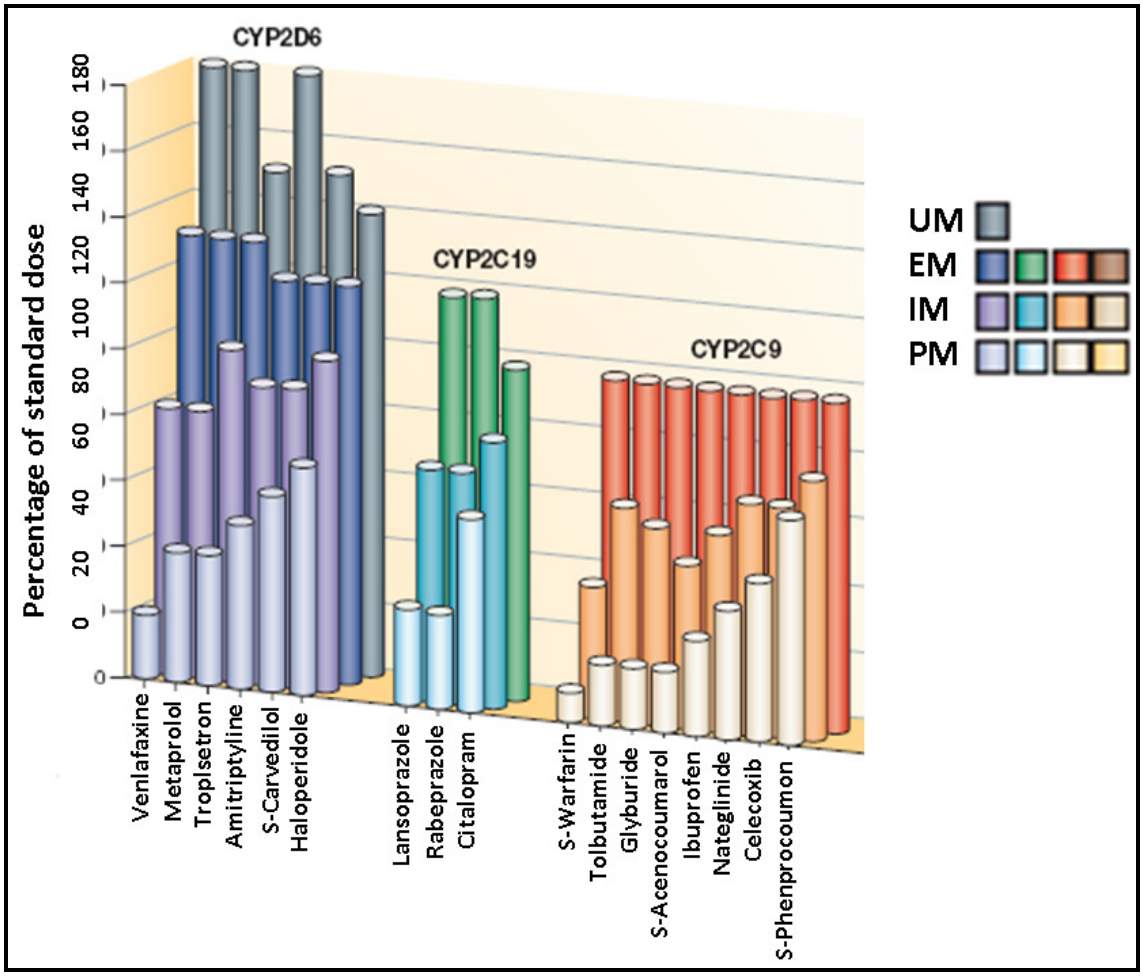


Figure 2.5: Percentage dosage adjustments for three main CYP genes.
(Kirchheiner *et al.* 2005) (Reprinted with permission from Nature Publishing Group)

2.2.2 The CYP2C Family

In the human genome 18 CYP families have been identified to date, of which the CYP2 family is the most diverse (Lewis 2004). The genes coding for the CYP2C enzymes in this family occur together in a

gene cluster found on chromosome 10q24.1-10q24.3 in the order *CYP2C8-CYP2C9-CYP2C19-CYP2C18* (Gray *et al.* 1995). These enzymes are collectively involved in the metabolism of 20% of prescribed drugs (Goldstein 2001). It is important to note that cardiovascular drugs, antiretroviral (ARV) drugs, oral hypoglycaemic agents and non-steroidal inflammatory drugs, all of which are metabolised by the *CYP2C* genes (Bertz and Granneman 1997; Ferguson *et al.* 2002,2005; Llerena *et al.* 2003; Nakamoto *et al.* 2007), are most frequently implicated in ADRs (Mehta *et al.* 2007).

Although all four of the *CYP2C* genes share a large amount of sequence similarity (Ingelman-Sundberg *et al.* 1999; Nebert and Russell 2002), their substrate specificity varies substantially. A closer look at the *CYP2C* genes reveals that all four genes consist of 9 exons. Despite the close affiliation that the *CYP2C* genes have to one another, studies have shown that only *CYP2C9* and *CYP2C19* exhibit variants occurring at a significant frequency, which affect the metabolism of ingested drugs (http://www.cypalleles.ki.se/cyp2c19.htm; Wanwimolruk *et al.* 1998; Bathum *et al.* 1999; Dandara *et al.* 2001; Gaikovitch *et al.* 2003; Hoskins *et al.* 2003; Halling *et al.* 2005). Thus, these two genes are the main focus for pharmacogenetic application with regards to the *CYP2C* family.

Of further interest, despite a large amount of sequence similarity observed between the four enzymes, the quantity of each enzyme expressed in the liver shows a large amount of variation, with CYP2C8:CYP2C9:CYP2C19:CYP2C18 expression occurring in the ratio 35:60:4:1 (Goldstein *et al.* 1994). Additionally, the metabolism activity of these enzymes is increased through exposure to inducers, in an expression level dependant manner with the strength of induction ranked *CYP2C8>CYP2C9>CYP2C19* (Chen and Goldstein 2009). This is important when considering co-administration of inducer drugs, as this may result in a UM *CYP2C* phenotype. The 20 fold difference in hepatic expression levels between *CYP2C19* and *CYP2C9* is of particular interest, as these two genes share the highest sequence similarity of 88.8% in the 2 kb upstream from the start codon (Kawashima *et al.* 2006). By comparing the subtle differences in 5'-upstream areas, where important promoter architecture is located, we could perhaps learn important information about the differences in the transcription systems of these two genes. This information includes epigenetic aspects, as well as the effect of subtle differences in nucleotide sequence on the recruitment of transcription factors (refer to Figure 2.6 for transcription factor binding sites identified in the *CYP2C* genes).

It is important to bear in mind that the evolution of the genome may provide important clues as to which regions are of functional importance. By comparing the paralogues and orthologues of *CYP2C19* to each other, regions that have been conserved throughout species and families can be

highlighted as areas of interest. It stands to reason that those areas that are conserved throughout are more likely to show functionality than those that are not (Hardison 2000; Aparicio *et al.* 2002; Prabhakar *et al.* 2006). With a maximum of 5% of the genome exhibiting DNA sequences that have been conserved throughout the course of evolution, the regions which are of functional validity can be sifted out from the so called "junk regions" (Pheasant and Mattick 2007). While coding regions are often the obvious place to search for conserved regions, less obvious regions may be elucidated through comparative sequence analysis.

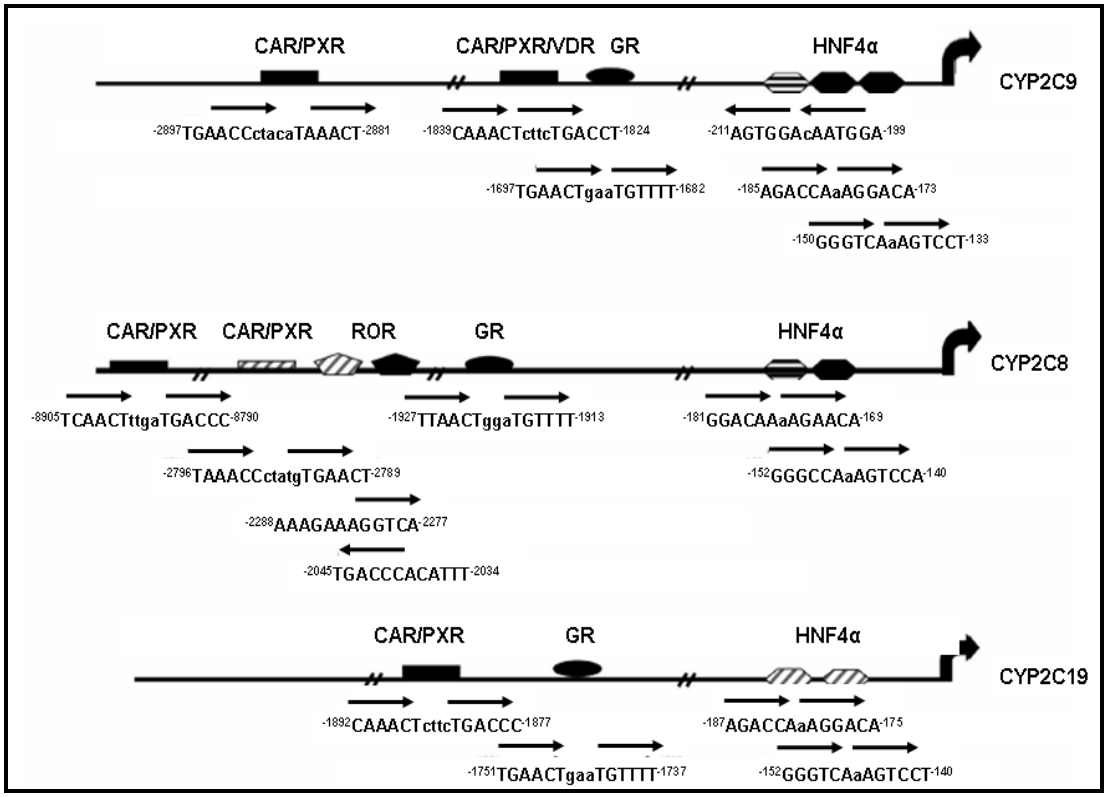


Figure 2.6: Transcription factor binding sites identified in the *CYP2C* genes through gel shift assays. (Chen and Goldstein 2009) (Reprinted with permission from Bentham Science Publishers)

2.3 *CYP2C19*

2.3.1 The *CYP2C19* Gene

In 1984 the genetic polymorphism responsible for the poor metabolism of S-mephenytoin was discovered. It was noticed that the deficient metabolism of this drug was inherited in an autosomal dominant fashion. After extensive research, an enzyme was identified which metabolised S-mephenytoin. This enzyme was CYP2C19 (Kupfer and Preisig 1984), the cloning of which was completed in 1994 by Goldstein *et al.* Today, individuals are phenotypically classified as PMs, IMs, EMs or UMs, by measuring the hydroxylation index in the urine after the ingestion of a standard dose of racemic mephenytoin (Goldstein and De Morais 1994). Although this method can be used in the clinical setting, the genotypic classification of individuals provides a much simpler method of

metaboliser status identification. It has been postulated that through the genotyping of *CYP2C19*, 93-100% of phenotypic PM metabolisers can be identified (De Morais *et al.* 1994a; Brosen *et al.* 1995; Chang *et al.* 1995; Kubota *et al.* 1996; Roh *et al.* 1996; Sagar *et al.* 1998). The detection of pharmacogenetic polymorphisms and the development of successful genotyping methods could have lasting consequences, as the cost of a genotyping test is less than the cost of a day in the hospital (Kirchheiner *et al.* 2001). It is therefore of vital importance that the variants present in *CYP2C19* are comprehensively studied in unique populations.

In line with traditional studies, most of the focus has been placed on the coding regions of *CYP2C19*; however it is important that the 5'-upstream, intronic and 3'-downstream regions are not neglected. Although they are not as extensively understood, their presence remains essential to gene functionality. In the promoter region of *CYP* genes, coding sequences for PXR, CAR, glucocorticoid receptor (GRE), hepatic nuclear factor (HNF)-3γ and HNF-4α have been implicated in the basal expression of the genes. Furthermore, it appears that additional HNF-4α increases the expression of *CYP2C9* and *CYP2C8*, but not of *CYP2C19* or *CYP2C18* (Gerbal-Chaloin *et al.* 2002). As all the genes contain DR1 elements, which have been shown to bind to HNF-4α, it is important to bear in mind the influence of other factors on the availability of binding sites when considering the differences they show with regards to the recruitment of transcriptional factors (Kawashima *et al.* 2006).

In the 5'-upstream region of the *CYP2C19* gene, CAR and GRE binding elements have been identified as far upstream as -1 891 bp, while a transcriptional repressor element has been identified as far downstream as exon 1 (Arefayene M *et al.* 2003; Chen *et al.* 2003). This further emphasizes the importance of studying beyond the 100 bp of the traditional core promoter region (Brown 2002). Furthermore, a recent study has identified a variant 806 bp upstream from the translational start site of *CYP2C19* which increases the transcriptional activity of the gene, thereby creating a class of UMs for CYP2C19 (Sim *et al.* 2006).

It is important to realise that while differences in drug response can vastly be attributed to genetic heterogeneity, the effect of epigenetics cannot be ignored. Where DNA variation may result in an altered gene product, epigenetics refers to the change in phenotype that, although linked to DNA, cannot be elucidated in terms of a change in the DNA sequence. Thus, epigenetics acts as the bridge between the environment and the genome (Ingelman-Sundberg *et al.* 2007). Epigenetics can refer to among other things, covalent modification of DNA and histones, DNA packaging, chromatin folding and regulatory noncoding RNAs (Gomez and Ingelman-Sundberg 2009). Epigenetic programming can alter in response to environmental stimuli such as drugs (Meaney and Szyf 2005), thus the study of epigenetics may be of particular interest to pharmacogenomic studies.

In the context of *CYP2C19* a CpG island has already been identified, which could impact the expression of the gene in certain individuals. Through methylation of the cytosines present in this island, transcription factors can either be blocked from binding, or enzymes responsible for chromatin remodeling may be recruited, subsequently resulting in a closed chromatin conformation, thus the expression of CYP2C19 will decrease (Ingelman-Sundberg *et al.* 2007). It has also been shown that environmental influences such as smoking can decrease the methylation status of CYP1A1 and therefore increase the expression (Anttila *et al.* 2003). The effect of the environment on methylation status in combination with the presence of the CpG island in *CYP2C19*, as well as the search for other putative islands, should thus be considered in order to obtain a comprehensive pharmacogenetic profile for *CYP2C19*.

With regards to the intronic regions, it is important to remember that intronic splice site mutations are not limited to the exon-intron boundaries, but that mutations further into the introns or within the exons themselves could play an equally important role. An excellent example of these splice site mutations is given by the most frequent null allele variant present in *CYP2C19*. This variant occurs in exon 5 of the gene and creates a new, stronger acceptor splice site (De Morais *et al.* 1994a). It is important to bear in mind that we are a long way from understanding all the elements involved in transcription and splicing, therefore we should keep an open mind before dismissing variants as having no effect on the gene. The genome should be viewed not as linear, but as a complex 3D structure, with every base pair playing a possible role in the tightening, loosening or shaping of its dynamic form.

2.3.2 *CYP2C19* Allele Nomenclature

Several different *CYP2C19* alleles have been described to date. Like other *CYP* genes, the functional copy of *CYP2C19* has been designated *CYP2C19*1.* Any other alleles containing variants affecting the enzyme product have been named *CYP2C19*2* all the way to *CYP2C19*26* on the Human *CYP* Allele Nomenclature website (http://www.cypalleles.ki.se/cyp2c19.htm) (refer to Appendix 1). Among these alleles, four null alleles have been identified to date, namely *CYP2C19*2*, *CYP2C19*3*, *CYP2C19*4* and *CYP2C19*7*. Both *CYP2C19*2* and *CYP2C19*7* are characterised by splice site mutations. *CYP2C19*2* creates a cryptic splice site in exon 5 as a result of a G>A change at position 681, while *CYP2C19*7* is found in intron 5 creating a mutation in the donor site (De Morais *et al.* 1994a; Ibeanu *et al.* 1999). The *CYP2C19*2* is the most common PM variant with the shift in the reading frame resulting in a premature stop codon. This premature stop codon creates a truncated enzyme product which lacks the heme binding region and is thus catalytically inactive (De Morais *et al.* 1994a). Similarly, *CYP2C19*3* is characterised by a G>A change at base position 636, which results

once again in a premature stop codon and truncated product (De Morais *et al.* 1994b). Finally the A>G change at position 1 which defines the *CYP2C19*4* allele, results in a GTG initiation codon, which greatly decreases the transcription/translation process of CYP2C19 (Ferguson *et al.* 1998).

Furthermore, several other alleles have been described with decreased or unknown effect on enzyme functioning, which may provide useful information regarding gene-based dosage recommendations, after thorough characterisation of these alleles. With regards to UM alleles, the *CYP2C19*17* allele has been shown to increase gene expression and thus enzyme activity, which in turn results in an increase in the metabolism of the prescribed drugs. Studies have shown that homozygous individuals for *CYP2C19*17* have a 2 times and 1.2 times lower metabolic ratio (MR) for omeprazole than wild type and heterozygous individuals, respectively. Similarly, these individuals have a 4.3 times and 3.7 times lower MR for mephenytoin (Sim *et al.* 2006). Electrophoretic mobility shift assay (EMSA) studies have shown that human hepatic nuclear factors bind to the -806 T variant, thus increasing the transcriptional activity. These studies were further validated by *in vivo* luciferase reporter transfection experiments performed in mice, which showed a two fold increase for the *CYP2C19*17* allele in comparison to the *CYP2C19*1* allele (Sim *et al.* 2006). The effect of the variant on transcriptional activity is depicted in Figure 2.7. The discovery of this variant in combination with significant frequencies of this variant in populations studied to date (refer to Table2.1), further emphasize the need to study the 5'-upstream region of this gene.

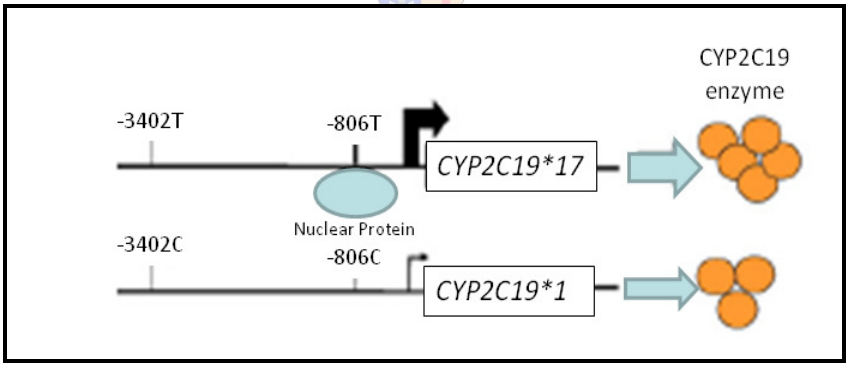


Figure 2.7: The effect of the -806T *CYP2C19*17* variant of the transcriptional activity of *CYP2C19*.
(Ingelman-Sundberg *et al.* 2007) (Reprinted with permission from Elsevier Limited)

The *CYP2C19*2* and *CYP2C19*3* alleles reported on the Human *CYP* Allele Nomenclature website, have been well studied and have been shown to occur at significantly different frequencies in different populations (Sistonen *et al.* 2009). Therefore, the implementation of population genetics to identify which alleles are present at what frequency in a specific population is an essential step in the reduction of ADRs in the population of interest. This has to a great extent been successfully implemented in most populations; however, to date African populations have been poorly

represented. Table 2.1 and Figure 2.8 provide an overview of the frequencies of the most important *CYP2C19* alleles identified to date in various populations. However data on some areas of the world, including Southern Africa, remain limited. To date *CYP2C19*17* has been inadequately studied as most of the studies were performed before this allele was identified. This is of importance to these studies as individuals designated with *CYP2C19*1* functional alleles may in actual fact exhibit *CYP2C19*17* or other alleles (Ragia *et al.* 2009). Thus, populations studied in this manner require re-evaluation.

Table 2.1: Allele frequencies of *CYP2C19*17* in different population groups

Population	Frequency of *CYP2C19*17 allele*	Reference
Chinese	0.04	Sim *et al.* 2006
Japanese	0.02	Sugimoto *et al.* 2008
Caucasians	0.18-0.25	Sim *et al.* 2006; Justenhoven *et al.* 2009; Ragia *et al.* 2009
Ethiopians	0.18	Sim *et al.* 2006

Figure 2.8: Distribution of *CYP2C19*2* and *CYP2C19*3* alleles throughout the world. (Sistonen *et al.* 2009) (Reprinted with permission from Wolters Kluwer Health)

2.3.3 Drugs Metabolised by CYP2C19

CYP2C19 is responsible for the metabolism of several clinically important drugs, including antidepressants, anticonvulsants, antiulcer agents, sedatives and antimalarial agents. CYP2C19 mainly detoxifies drugs; however, it is also responsible for converting certain pro-drugs, such as proguanil and chloroproguanil to active molecules (Bertilsson *et al.* 1989; Helsby *et al.* 1990; Wan *et al.* 1996; Khaliq *et al.* 2000). When reviewing recent studies examining *CYP2C19* genotype and ADR

associations, two of the most promising associations appear to point towards clopidogrel, an anti-platelet agent and tamoxifen, an anti-estrogen agent. *CYP2C19* UMs have been shown to respond better to tamoxifen treatment (Schroth *et al.* 2007), while *CYP2C19* PMs are less likely to respond to clopidogrel treatment and are more likely to experience ADRs such as a cardiovascular ischemic event or even death (Shuldiner *et al.* 2009). As a result, the FDA have recently changed the prescribing information for clopidogrel to include the impact of *CYP2C19* genotype (Ellis *et al.* 2009).

With regards to the metabolism of antidepressants, CYP2C19 is involved in the metabolism of moclobemide, amitriptyline, clomipramidine, sertraline and citalopram, which have been shown to exhibit associations with ADRs and drug plasma concentrations (Schweizer *et al.* 2001; Yokono *et al.* 2001; Yu *et al.* 2001; Herrlin *et al.* 2003; Kirchheiner *et al.* 2005; Steimer *et al.* 2005). It is important to note here that major depressive disorder is among one of the leading causes of death and disability worldwide (Murray and Lopez 1997). Furthermore, it has been documented that 30-50% of patients on antidepressants do not respond to medication (Entsuah *et al.* 2001; Steimer *et al.* 2001; Bauer *et al.* 2002; Thase 2003), which may also result in a higher likelihood of these patients committing suicide (Zackrisson *et al.* 2009). Thus, it is crucial that the treatment of depression through antidepressants receives urgent attention.

Interestingly, in the case of proton pump inhibitor (PPI) metabolism, it is not necessarily a disadvantage to be a PM. Often it is the EMs in which treatment failure occurs (Gardiner and Begg 2006). PPIs are the most extensively used drug class in the world of which omeprazole, lansoproazole, pantoprazole and rabeprazole are metabolised by CYP2C19 (Andersson *et al.* 1998), with omeprazole falling into one of the 10 most prescribed drugs worldwide (Chen *et al.* 2003). Individuals exhibiting PM phenotypes display greater acid suppression with PPI treatment (Furuta *et al.* 1999; Sagar *et al.* 2000; Shirai *et al.* 2001) with better *Helicobacter pylori* cure rates of 84-92% and 98-100% in heterozygous and homozygous PMs respectively as opposed to 60-73% in EMs with omeprazole and lansoprazole treatment (Furuta *et al.* 2001; Sapone *et al.* 2003) (refer to Figure 2.9). By applying genotype information to treatment plans, cheaper, easier dual-therapy treatment can be used for the cure of *H. pylori* in PMs as opposed to the triple-dose therapy or a non- PPI alternative (Aoyama *et al.* 1999; Tanigawara *et al.* 1999). This being said, it has been documented that PM individuals who use long term omeprazole treatment (20 mg/day for more than a year) show decreased levels of vitamin B_{12} serum levels (Sagar *et al.* 1999) (refer to Figure 2.10). Therefore, as is the case in all treatment regimes, it is important that all available information is reviewed before treatment is applied.

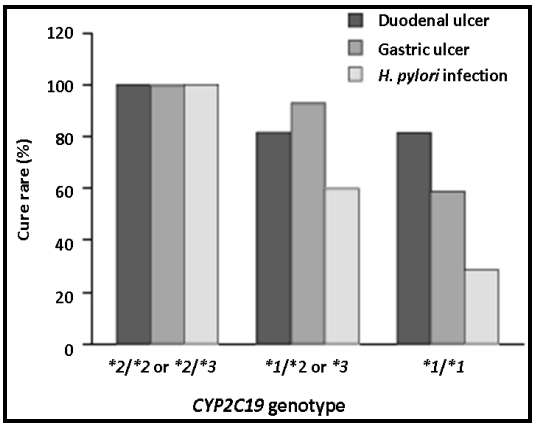


Figure 2.9: Cure and healing rates for *H. pylori* infection, gastric and duodenal ulcers for PM, IM and EM individuals after treatment with 20 mg/day omeprazole for 2 weeks.

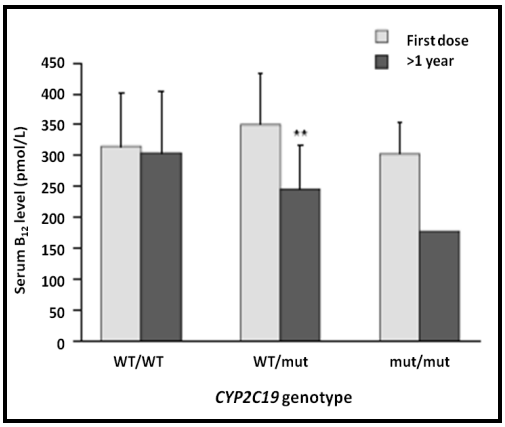


Figure 2.10: Vitamin B_{12} serum levels or EM, IM and PM individuals after an omeprazole treatment of 20 mg/day for one day and for more than a year.

(Furuta *et al.* 1998; Sagar *et al.* 1999). (Reprinted with permission from Wolters Kluwer Pharma Solutions)

Specifically, when considering African populations, the metabolism of anti-malarial drugs such as proguanil and anti-HIV agents such nelfinavir by CYP2C19 is important, as both malaria and HIV/AIDS are predominant in Africa. With regards to proguanil, clear associations have been made between *CYP2C19*17* and the drug plasma concentration (Janha *et al.* 2009; Kerb *et al.* 2009). Considering that there are reported to be 300 million cases of malaria every year, of which one million result in death, 90% of which occur in Africa, with costs related to malaria amounting to $2 billion a year (http://www.malaria.org.za/), it is essential that the treatment of malaria is at an optimal level. Furthermore, with regards to nelfinavar, a high plasma concentration has been found in PMs (Haas *et al.* 2005). According to the world health organization (WHO) global summary of the AIDS epidemic, December 2007 (http://www.who.int/hiv/data/2008global_summary_AIDS_ep.png), 33 million people are living with HIV, of which 2.7 million were infected and two million died in 2007. In South Africa alone, 5 700 000 people were living with HIV in 2007, of which 350 000 died (http://www.who.int/GlobalAtlas/predefinedReports/EFS2008/index.asp?strSelectedCountry=ZA). Thus, as CYP2C19 may be involved with both anti-malarial and anti-HIV agent metabolism, the elucidation of *CYP2C19* genotypes in African populations is a valuable research avenue. (For a list of drugs affected by CYP2C19 genotype status, refer to Table 2.2).

Table 2.2: CYP2C19 drug response according to metaboliser status

Drug	Drug type	Response	Reference
Omeprazole	PPI	Treatment failure in EMs Reduced serum vitamin B12 in PMs	Xie *et al.* 1999; Sagar *et al.* 1999
Citalopram	Antidepressant	GI disturbances and restlessness	Kirchheiner 2001; Herrlin *et al.* 2003;
Sertraline	Antidepressant	GI disturbances, dizziness	Wang *et al.* 2001
Diazepam	Sedative	Longer to emerge from anaesthesia and unconsciousness	Goldstein *et al.* 2001; Inomata *et al.* 2005
Flunitrazepam	Sedative	Sedation[1]; Psychomotor impairment	Kilicarslan *et al.* 2001
Phenobarbital	Sedative	Neurotoxicity	Gidal and Zupanc 1994
Terodiline	Anticonvulsant	Cardiotoxicity	Ford *et al.* 2000
Hexobarbital	Anticonvulsant	Sedation[1]	Knodell *et al.* 1988
Mephenytoin	Anticonvulsant	Sedation[1]	Kupfer *et al.* 1984
Clarisoprodal	Muscle relaxant	Drowsiness[1]	Olsen *et al.* 1994
Proguanil	Antimalarial agent	Gastrointestinal ADR Malarial prophylaxis	Kaneko *et al.* 1999 Gardiner and Begg 2006
Chloroproguanil	Antimalarial agent	Gastrointestinal ADR	Kaneko *et al.* 1999
Voriconazole	Antifungal	Hepatotoxicity	Boucher *et al.* 2004
Tamoxifen	Anti-cancer	Carrier UMs more likely to benefit and decreased breast cancer risk	Schroth *et al.* 2007; Justenhoven *et al.* 2009
Cyclophosphamide	Anti-cancer	Poor renal response in PMs PM and IMs show a lower risk of developing premature ovarian failure	Takada *et al.* 2004
Clopidogrel	Antithrombotic	Poor reduction of platelet aggregation in IM and PMs Increased risk of cardiovascular ischemic event or death in PMs or IMs	Hulot el al 2006 Shuldiner *et al.* 2009
Nelfinavor	Anti-HIV agent	PMs respond favourably due to increased virologic response	Haas *et al.* 2005

[1]It is important to note that sedation and drowsiness have more serious implications to general functioning such as impaired driving and that less serious ADRs may affect compliance

For guidelines in relation to the application of pharmacogenetic data, studies have already been performed to recommend drug dosages according to genotype. These guidelines suggested a 60% and 110% decrease or increase in the standard drug dosage for PM and EM individuals respectively (Kirchheiner *et al.* 2004). The adjustment in dosage according to genotype is represented in Figure 2.11. It is important to realise that CYP2C19 does not always act alone in the metabolism of toxins and is often complemented by the action of CYP2C9, CYP2D6 and CYP3A4 (Gardiner and Begg 2006), thus concurrent genotyping of all relevant pharmacogenetic biomarkers would be the eventual goal of pharmacogenetics. Although CYP2C19 is not the only enzyme responsible for the metabolism of these drugs, it remains important to identify and understand the different *CYP2C19* alleles.

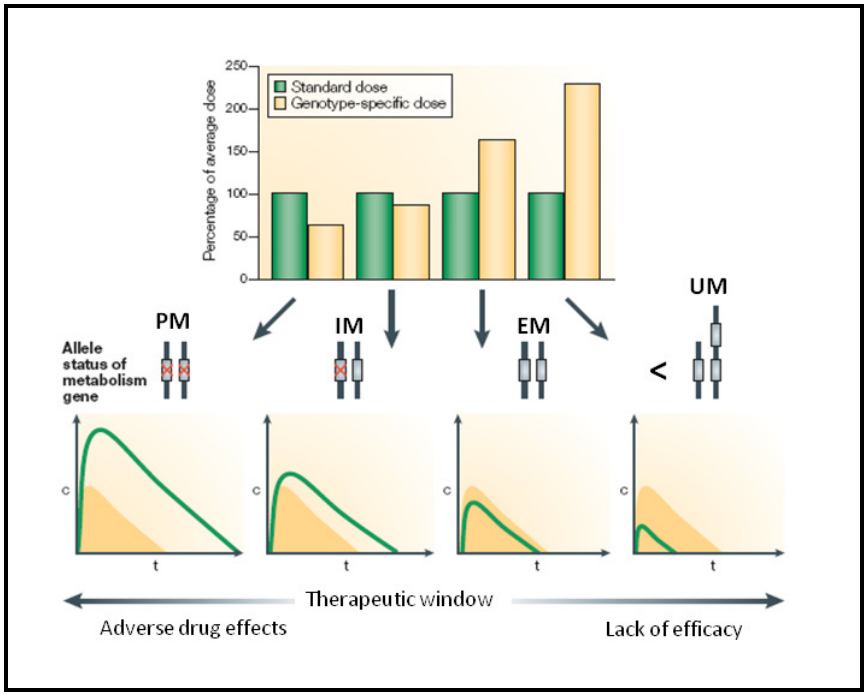


Figure 2.11: Dosage adjustment according to genotype.
(Kirchheiner *et al.* 2005) (Reprinted with permission from Nature Publishing Group)

2.4 Pharmacogenetic Applications and Success Stories

There have already been a few success stories in the world of pharmacogenetics, further emphasizing the need for the advancement and acceptance of the field and reiterating the viability of such diagnostic tests. One story tells of a patient with heart problems who was prescribed warfarin. This patient underwent five years of trial and error dosage testing and risk of haemorrhaging. In these five years, 20 samples of blood were taken from the patient in the hope of determining the appropriate warfarin dosage. Finally the patient was genotyped for *CYP2C9* and subsequently identified as a PM, resulting in the establishment of the correct warfarin dosage (Flockhart and McMillin 2006). Initial genotyping of *CYP2C9* could have decreased time, cost and suffering for both the patient and the health care institute.

Despite the fact that Tykerb, a cancer treatment drug, is mainly metabolised by CYP3A4 and CYP3A5, with CYP2C19 only playing a minor role in metabolism, it was found that CYP2C19 genotype was associated with ADRs related to the drug. It was observed that 15% of the patients showed rash and diarrhoea side effects. The only association that was found with these side effects was with *CYP2C19* variants. Furthermore, all patients homozygous for *CYP2C19*2* showed the side effects (Nelson and Dolder 2006). Thus from this example we can see that although the enzyme of focus may not be involved in the main pathway of metabolism, the role that it plays cannot be ignored.

In an example highlighting the need for population specific directed studies, the implementation of BiDil, a drug used in the treatment of severe heart failure within the African American population, decreased the death and hospitalisation of this population significantly. This same beneficial response to the drug was, however, not observed in other populations (Meadows 2005). Although the understanding of human genetic diversity between different population groups may be a valuable resource, it is important that these differences are celebrated rather than abused. It is essential that we bear in mind that the genetic diversity observed is not exclusively present between populations, but also occurs within population groups and that it is this diversity that is responsible for enriching the *Homo sapiens* species (Lahn and Ebenstein 2009).

Although success stories for pharmacogenetics have been documented, most pharmaceutical companies cannot perceive the benefits of genotyping and the one-drug-fits-all profile seems more profitable. However, let us take into account that the release of one new drug costs a company approximately R1 billion and takes about 12 years to reach the release stages (Munos 2006). Additionally, only one in every 12 drugs is actually released to the public (Willard and Ginsburg 2009) and out of the 499 drugs approved since 1980, 21% have had to change their dosage recommendations post launch, of which 79% of these changes were due to safety related issues (Cross *et al.* 2002). By considering genotype before drugs are designed and released, pharmaceutical companies could potentially save a substantial amount of time and money.

The pharmaceutical company, Amgen, has begun to realise the benefits of obtaining genomic information. Initially Amgen released a colon cancer drug, Vertibix, with an overall effectiveness of only 10% and the drug was thus rejected. After the implementation of a genotyping assay supplied by DxS, it was discovered that it was in fact patients with KRAS mutations who were resistant to the drug. When these patients were removed from the treatment pool, the response rate increased dramatically and the drug was accepted (Allison 2008). Amgen has since begun the Women's Genome Health Study in which samples from 28 000 healthy women have been collected and followed for more than a decade. With the use of Illumina's genotyping platform, a genome wide

scan of 360 000 variants is currently being performed for these women. Amgen expects to reap the benefits of this study in the near future (Allison 2008). It is expected that in the future more pharmaceutical companies will incorporate pharmacogenetic considerations when developing drugs, therefore emphasising the importance of studies such as this one.

2.5 The South African Context

2.5.1 Health Care in South Africa

South Africa is a developing country, providing opportunities for building and improving the health care systems in South Africa, however, the under-staffed and under-equipped public health care institutions leave much to be desired. The lack of appropriate health care has severe consequences, of which ADRs and treatment failure are highly significant (Mehta *et al.* 2007). A study conducted in a hospital in the Cape Town metropolitan area, to determine the incidence of ADRs in a South African context, showed that 14% of all patients admitted to the hospital of interest exhibited ADRs (Mehta *et al.* 2007). Additionally, the fatality rate observed was 5-10 fold higher than reported in USA and UK populations. More than half of the ADRs led to hospitalisation and almost a third of hospital-acquired ADRs were preventable. In the already under-equipped health care system of South Africa, these preventable hospitalisations need to be eliminated. The ADRs observed in this study included those affecting metabolic, renal, hepatobiliary, neurological, haematological, endocrine, skin, mucosa, cardiovascular, gastrointestinal, immune, respiratory and musculoskeletal systems. A large percentage of these ADRs were related to the prevalence of HIV/AIDS and antiretroviral therapy (ART) in the country. The study noted that in patients younger than 60 years, those infected with the HIV virus were twice as likely to develop ADRs, while individuals receiving ARTs were 10 times more likely to develop ADRs than those not taking ARTs. It is thus important that the HIV and ART status of the patient is taken into account when considering the treatment of patients in the South African context.

Considering that HIV/AIDS, depression and malaria are all treated with drugs metabolised by CYP2C19, it is important to consider the prevalence of these diseases in the South African context. In a recent national population census, it was found that 0.6% of South Africans suffered from severe mental illness and that mental and nervous system disorders were secondary only to HIV/AIDS (Seedat *et al.* 2004) which occurs at a frequency of 11.6% in the population (http://www.unaids.org/en/Regions_Countries/Countries/south_africa.asp). Additionally, as 90% of the deaths associated with malaria occur in sub-Saharan Africa (http://www.malaria.org.za/); the treatment of malaria is an important consideration for South Africa. By determining the common genotypes of genes with pharmacogenetic applications in South African populations, the correct

treatment plans can be implemented and optimised. These treatment plans should include the elimination of ADRs and treatment failure with the aid of easy, efficient genotyping protocols.

The substantial wealth of information available regarding genes with pharmacogenetic applications should compliment studies on these genes in South African populations. Information regarding mutations in the genes and the interactions which occur between the genes will hopefully play an important role in improving South African health care through the partial elimination of ADRs and treatment failure. Although there is resistance to the introduction of genetic applications in the general clinical setting, this is mainly due to a lack of education. When calculated in simple terms, the benefits become clear. In terms of PPI treatment, a homozygous EM needs treatment for three months, a heterozygote for two months and a PM for only one month (Furuta *et al.* 1998); thus if the heterozygotes and PMs are identified prior to treatment, one month and two months of treatment could be eliminated respectively. The cost of genotyping these individuals is less than the cost of the unnecessary treatment. In summary ADRs and treatment failure are a significant consequence of medication utilised within South Africa and the genotyping of C*YP2C19* could aid in the elimination of these affects.

2.5.2 The Rainbow Nation

The South African population can broadly be divided into four major groups, namely the black African (79.2%), Caucasian (9.58%), Mixed Ancestry (8.91%) and Indian and Asian (2.49%) populations (http://www.statssa.gov.za/census01/HTML/default.asp). These population groups can subsequently be further subdivided into 11 official linguistic groups (http://www.statssa.gov.za/census01/HTML/default.asp). A closer look at each of these population groups reveals likely divergences between the different groups, due to their different ancestry. With regards to the Caucasian populations, it has been suggested that the Afrikaans-speaking group provides a textbook example of a founder population (Ridley, 2004); therefore this group may yield novel results and warrants closer inspection. The mixed ancestry population on the other hand have been reported to exhibit the highest level of admixture worldwide (Tishkoff *et al.* 2009); therefore this population may exhibit a unique genetic profile, when compared to the other South African populations. Lastly the genetic substructure observed between the African populations residing within South Africa merits further inspection. Figure 2.12 gives an indication of the relatedness between the Eastern Bantu-speaking populations residing in South Africa (Lane *et al.* 2002).

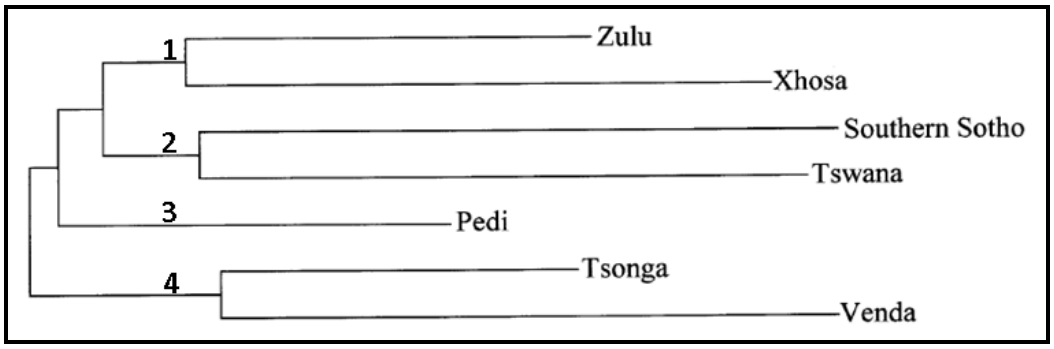


Figure 2.12: The genetic substructure of the Eastern Bantu-speaking populations of South Africa, according to Y-chromosomal data.
(Lane *et al.* 2002) (Reprinted with permission from John Wiley and Sons)

With regards to *CYP2C19*, only the Venda population has, to our knowledge, previously been examined in South Africa (Dandara *et al.* 2001; Matimba *et al.* 2009). It would, therefore, be advantageous to examine other larger South African groups to determine to what extent the populations differ from one another. When referring to Figure 2.12, each branch, numbered 1-4, represents a different population cluster. Thus, in order to accurately determine the relatedness between the different populations, at least one population from each cluster should be examined.

As the Venda population (on branch 4) has already been analysed with regards to *CYP2C19*, this study aims to provide data for *CYP2C19* in the Xhosa population (on branch 1). The data obtained from this study can be added to the data available for other genes with pharmacogenetic application such as *CYP2D6* (Wright *et al.* 2009, manuscript submitted to *Prog Neuropsychopharmacol Biol Psychiatry*) and *CYP3A4* (Warnich *et al.* 2009, manuscript submitted to *Prog Neuropsychopharmacol Biol Psychiatry*) in the Xhosa population and subsequently compared to already existing data on *CYP2C19* in other populations. Data for *CYP2C19* genotype in African populations can be viewed in Table 2.3 and Table 2.4, however, the comparisons are unfortunately complicated by small sample sizes (refer to Table 2.3) and a lack of complete genotype data (depicted by the shaded blocks in Table 2.3 and Table 2.4 where data for these variants are not available). This emphasises once again the need for comprehensive genotyping platforms to identify novel and lesser characterised variants in the complex African populations. This study aims to contribute towards the achievement of this goal.

Table 2.3: Allele frequencies detected via the sequencing of *CYP2C19* in different African populations (Matimba *et al.* 2009)

	Hausa (n = 20)	**Yoruba (n = 20)**	**Ibo (n = 20)**	**Luo (n = 30)**	**Maasai (n = 13)**	**Shona (n = 15)**	**Venda (n = 9)**	**Tanzanian Bantu (n = 10)**
***17**								
***15**	0	0	0	0	0	0	0	0.05
E92D	0	0	0	0	0	0.03	0	0
V113I	0	0	0	0	0	0	0.06	0
***9**	0	0	0	0	0	0	0	0
***22**	0	0	0	0	0	0	0	0.06
***3**	0	0	0	0	0.04	0	0	0
***2**	0.13	0.15	0.33	0.07	0.08	0.23	0.17	0.15
V331I	0.03	0.03	0	0	0.04	0	0	0
***13**	0	0	0.04	0.03	0	0.03	0	0
***12**	0	0	0	0.04	0	0.03	0	0

Table 2.4: Allele frequencies in different African populations detected through RFLP genotyping

	African American (n = 481) de Leon *et al.* 2009	**Egyptian (n = 247)** Hamdy *et al.* 2002	**Zimbabwean (n = 76)** Dandara *et al.* 2001	**Ethiopian (n = 114)** Persson *et al.* 1996	**Ugandan (n = 99)** Miura *et al.* 2008
***17**				0.140	0.340
***15**					
E92D					
V113I					
***9**					
***22**					
***3**	0.001	0.002	0	0.020	0.020
***2**	0.183	0.110	0.131	0.140	0.250
V331I					
***13**					
***12**					

2.6 Aim of the Study

Hypothesis: Genetic variation of the *CYP2C19* gene in the Xhosa population differs from other populations, including those of African ancestry. Therefore, for pharmacogenetic applications, the screening tests for individuals from this population should be designed to fit the genetic profile of this specific population.

Objectives:

- To identify the spectrum of common sequence variation present in the *CYP2C19* gene in the Xhosa population, semi-automated sequencing of 15 Xhosa individuals was performed.
- To determine the effect of novel variants, various *in silico* analyses were performed.
- To confirm the frequencies of the most important variants identified, restriction fragment length polymorphism (RFLP) analysis was performed in an additional 85 Xhosa individuals.
- To determine *CYP2C19* CNV, duplex real-time PCR was performed in the entire cohort of 100 healthy Xhosa individuals.
- The data obtained from the study were compared to the data obtained in previous studies and from this and other data, educated genotyping protocols can be devised to fit the genetic profile of the Xhosa population.
- To validate the results obtained from the *in silico* analyses for the detected 5'-upstream variants; dual luciferase reporter assays were performed as functional validation.
- To investigate lesser studied genomic regions, multi-species comparative sequence analyses were utilised to identify conserved regions located in the *CYP2C19* 5'-upstream region
- CpG island analysis was utilised to identify putative CpG islands which may be involved in the epigenetic control of the *CYP2C* genes.

After analysis of these data, the generated genetic profiles for *CYP2C19* aim to aid in the genotyping of other South African populations and thereby facilitate optimised treatment plans in the South African context.

CHAPTER 3:

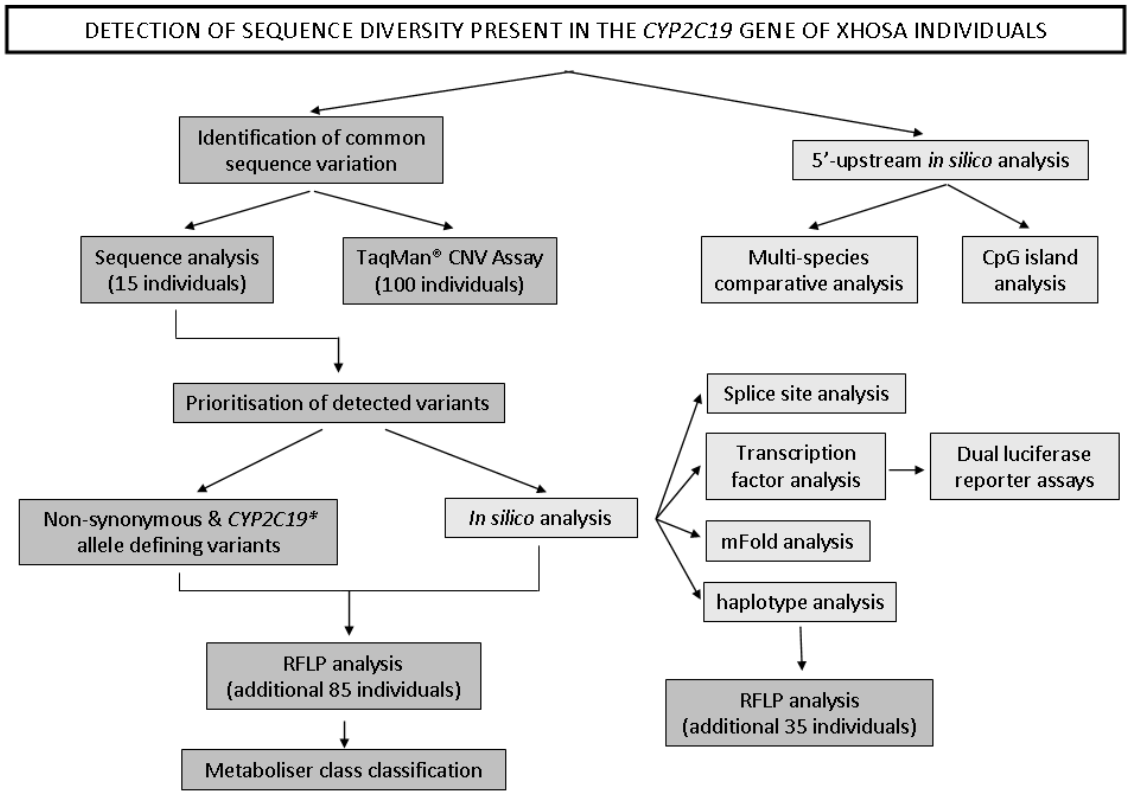
MATERIALS AND METHODS

CHAPTER 3: MATERIALS AND METHODS

3.1 Patient Samples

This study, which forms part of a larger pharmacogenetics project, was granted institutional approval by the Stellenbosch University Research Ethics Committee (REC) in terms of the Health Act No 61.2003. For DNA collection purposes, written, informed consent was obtained from 500 schizophrenia and 500 control individuals from the Xhosa population residing in the Western Cape (refer to Appendix 2). As selection criteria for Xhosa ethnicity, all four of the participants' grandparents were required to be of Xhosa origin. Prior to this study, genomic DNA (gDNA) was extracted from the venous blood obtained from the participants using the Miller *et al.* (1988) protocol (refer to Appendix 3.1). The quality, purity and concentration of each DNA sample was determined using a spectrometer (NanoDrop® ND-100, Nanodrop Technologies Inc., Wilmington, Delaware, USA), measuring the absorbance of the gDNA at 260 nm. For this study gDNA from 100 of the Xhosa control individuals was used. Fifteen of these samples were selected, as they have been sequenced for other *CYP* genes (Warnich *et al.* 2009, manuscript in preparation) and utilised for semi-automated sequence analysis in order to ensure that alleles occurring at a frequency of 10% will be detected with 95.8% certainty (from the binomial distribution).

3.2 Strategy of Study



3.3 Screening for Variation in the Xhosa Population

3.3.1 Primer Design

The *CYP2C19* DNA sequence was obtained from the National Centre for Biotechnology Information (NCBI) Entrez core nucleotide sequence (reference number NM 000769) and Ensembl website (http://www.ensembl.org/index.html) (Ensembl Gene ID ENSG00000165841). Primers were designed from these sequences using OligoAnalyzer 3.0 (http://www.idtdna.com/ analyzer/Applications/OligoAnalyzer/) and Primer3 (Rozen and Skaletsky 2000) computational tools. Primers were designed to allow for the amplification of all nine exons including exon-intron boundaries, the 346 bp of the 3' untranslated region (UTR), first 1.03 kb of the 5' UTR as well as an upstream 5'UTR region ranging from 1 552 bp to 2 095 bp upstream from the translational start site, selected according to functional evidence in the literature (Chen *et al.* 2003). All primers that were used to amplify from gDNA were designed to avoid co-amplification of isoforms, thus primers were also designed for nested PCR reactions as required. The primer sequences are shown in Table 3.1.

3.3.2 Polymerase Chain Reaction (PCR) Amplification

All PCR amplification reactions were prepared to a final volume of 25 µl, containing a final concentration of 1X buffer and 0.4 mM dNTPs. All reaction mixes utilised 15 ng of gDNA, with the exception of the nested PCR reactions, which used a 1/100 dilution of PCR product. Excluding the reaction mix for the amplicon containing exon two and three (E2+3), which utilised 0.2 µM of each primer and 1 U of Taq, the reaction mixes all contained 0.4 µM of each primer and 0.5 U of Taq. All amplification cycle reactions were performed at an initial denaturation of 94°C for 3 min; followed by varying cycles of denaturation at 94°C for 15 sec, annealing for 15 sec and extension at 72°C. All cycles were concluded with a final extension step at 72°C for 5 min. The specifications of each PCR reaction, the sizes of the amplicons and the suppliers of the reagents utilised are indicated in Table 3.2.

3.3.3 PCR Product Visualisation

The resulting PCR amplicons were analysed using gel electrophoresis. Visualisation of the fragments was executed by loading 5 µl of cresol loading buffer, mixed with 5 µl PCR product onto 1% (w/v) ethidium bromide-stained agarose gels. Electrophoresis was performed at 120 V for 30-45 min depending on the size of the fragments. The size of the amplicons was determined using Hyperladder IV (Bioline, London, UK) as a 100 bp molecular weight marker for all amplicons smaller than 1 kb, while Hyperladder I (Bioline, London, UK) was used as a 1 kb molecular weight marker for amplicons exceeding 1 kb in size. After electrophoresis, the fragments were visualised under a UV light at A_{260} nm with a MultiGenius Bio Imaging Capture System (Syngene, Cambridge, UK).

Table 3.1: Primer sequences

Region	Template Primers (5' – 3')		Internal Primers (5' – 3')	
P+E1	CYP2C19P+E1:F	CAG AAC TGG AAC ACC TAG CTC TCA		
P+E1	CYP2C19P+E1:R	GAC AGA CTG GAA AAG GCA ACA AAA G		
P			CYP2C19P+E1:R1	CTC ACA CCT CAT ATC CCT TTG GAA TCT CTC
P			CYP2C19P+E1:F3	GTG TCT TCT GTT CTC AAA GCA TCT C
P			CYP2C19P+E1:FS1	GGA CAA AGT CTC CTA ATC TTC GAT ATA G
P			CYP2C19P+E1:RS1	CAC CGT CAT AAT TGA GAG CAC TGA AG
P			CYP2C19P+E1:mF2020	CTA AAG AGA GCA ACC AAG CT**g** AT
P			CYP2C19P+E1:mF*17	GTG TCT TCT GTT CTC Aa**t** G
E1			CYP2C19P+E1:mF*15	CTC TCA TGT TTG CTT CT**g** **a**TT TCA
E1			CYP2C19P+E1:mF*4	CTT AAC AAG AGG AGA AGG Ct**a** CA
E2+3	CYP2C19E2+3:F	CAT AAA AGA CTG TTG GAC CAG G		
E2+3	CYP2C19E2+3:R	AGG AGA GCA GTC CAG AAA GG		
E3			CYP2C19E2+3:FS1	GCA CAC CTA CCA AAT CCT CTG
E2			CYP2C19E2+3:RS1	GGA GAT CCC AGG CAA GAA AGA GG
E4	CYP2C19E4:F	GCA ACC ATT ATT TAA CCA GCT AGG		
E4	CYP2C19E4:R	TCA AAA ATG TAC TTC AGG GCT TGG TC		
E5	CYP2C19E5:F	CAA CCA GAG CTT GGC ATA TTG T		
E5	CYP2C19E5:R	GCA GAA CAG AGC TTT TCC TAT C		
E5			CYP2C19E5:mF*7	CTT CCT GAT CAA AAT GGA G**c**A **c**G
E5			CYP2C19E5:mF*10	GGT TTT TAA GTA ATT TGT Ta**c** **c**GG TTC C
E6	CYP2C19E6:F	GCA TTC CCT TTG AAA ACT GGC ACA AGA C		
E6	CYP2C19E6:R	CAC ACC ATT AAA TTG GGA CAG ATT ACA GC		
E7	CYP2C19E7:F	GGG CTT CTC TTC CTT CTT TCA TTT CT		
E7	CYP2C19E7:R	CTC TCA CCC AGT GAT GGT AGA GGG		
E8	CYP2C19E8:F	CGT CTA TCT GTC TGG AAA TGG		
E8	CYP2C19E8:R	GAG GAT GTA TCA CCA GCG		
E9	CYP2C19E9:F	CAC CCA TCC ATC CTT TCA TTC ATG C		
E9	CYP2C19E9:R	GGA CCA GAG GAA AGA GAG CTG	CYP2C19E9:mF	CAC ATG AGG AGT AAC TTC TCC **a**T
E9			CYP2C193UTR:F	CAC ATG AGG AGT AAC TTC TCC CT
3'UTR	CYP2C193UTR:R	CCT CAT GTA ACT CTA AAT TTT GG		

P: promoter, E: exon, 1-9: the specific exon numbers, F: forward primer, R: reverse primer, S: sequencing primer, m: mutagenic primer, bold lower case letter: mutagenic bases

Table 3.2: PCR amplification specifications

PCR reagents were supplied by either Bioline, London, UK or Kapa Biosystems, Cape Town, SA

Region	Products used	$MgCl_2$ conc (mM)	Number of cycles	Annealing temp (°C)	Extension time (sec)	Size (bp)
P+E1	Bioline	3.5	40	68	180	2 411
E2+3	Bioline	3	40	58	60	996
E4	Bioline	2	10;30	60;55	30	472
E5	Bioline	2.5	10;30	65;60	30	523
E6	Bioline	0.5	40	48	30	477
E7	Kapa	-	10;30	55;50	30	562
E8	Kapa	-	10;30	55;50	30	452
E9	Kapa	-	10;30	60;55	30	418
3'UTR	Bioline	2	40	62	60	742
Nested	Bioline	2	25	55	30	

P: Promoter, E: Exon, Nested: nested PCR

3.3.4 Sequence Analysis

For optimal sequencing results, nested PCRs were performed on fragments larger than 1 kb in size and internal sequencing primers were designed at 700 bp intervals to allow for efficient bi-directional semi-automated sequencing. Amplicons were purified with the use of Sure Clean (Bioline, London, UK) according to the specified manufacturer's protocol (refer to Appendix 3.2). The concentration of the purified product was subsequently measured with a spectrometer (NanoDrop® ND-100, Nanodrop Technologies Inc., Wilmington, Delaware, USA) by measuring the absorbance of the amplified DNA at 260 nm. The amplicons of 500-1 000 bp were subsequently diluted to 20 ng/µl, while those of 200-500 bp were diluted to 10 ng/µl. Sequencing reactions were performed on these diluted products with the utilisation of Big Dye v3.1 sequencing chemistry, accompanied with the addition of Half Dye mix (Bioline, London, UK) according to the manufacturer's recommendations (refer to Appendix 3.3). All sequencing cycle reactions were performed at an initial denaturation of 94°C for 5 min, followed by 25 cycles of denaturation at 94°C for 10 min, annealing at 55°C for 10 min and extension at 60°C for 4 min. Subsequent purification was performed via the addition of 0.2% SDS at a purification cycle of 98°C for 5 min and 25°C for 10 min. Capillary electrophoresis was performed by the Central Analytical Facility of Stellenbosch University on a 3130Xl Genetic Analyzer, according to the manufacturer's protocol (Applied Biosystems™, Foster City, California, USA).

3.3.5 Identification of Variants

The data obtained from the 15 sequenced Xhosa DNA samples were examined and compared to the *CYP2C19* reference sequence, ENSG00000165841 (see Section 3.2.1), using Sequencher™ v4.7 Build 2946 Demo Version (Gene Codes Corporation, 2006), BioEdit v7.0.9.0 (Hall 1999) and SeqScanner v1.0 (Applied Biosystems™, Foster City, California, USA) to identify variants present in the sequenced samples. To ensure consistency with the international Human *CYP* Allele Nomenclature website

(http://www.cypalleles.ki.se/), all nucleotide bases were named according to this system, with the first coding nucleotide annotated as +1 and the nucleotide directly upstream as -1.

3.3.6 Prioritisation of Variants

The variants identified were prioritized for genotyping in an additional 85 healthy Xhosa individuals according to the following criteria: i) non-synonymous mutations, ii) functionality according to literature-based evidence obtained from the Human CYP Allele Nomenclature website, iii) *in silico* splice site, promoter and mFold analysis. In addition to the variants prioritised in this manner, those variants not described on the Human *CYP* Allele Nomenclature website, but occurring in perfect linkage disequilibrium (LD) with known variants, were genotyped in a total of 50 healthy Xhosa individuals, including the 15 sequenced individuals, to determine whether they remained in perfect LD. Furthermore, those variants not observed in the 15 sequenced samples, but documented to abolish CYP2C19 enzyme activity, were genotyped in the entire cohort. Lastly, additional genotyping protocols were devised to detect the non-functional *CYP2C19*3*, *CYP2C19*4* and *CYP2C19*7* variants as well as to ensure that all identified *CYP2C19*2* (rs4244285) variants were not erroneously genotyped as a result of the adjacent *CYP2C19*10* (rs6413438) variant (refer to Figure 3.1).

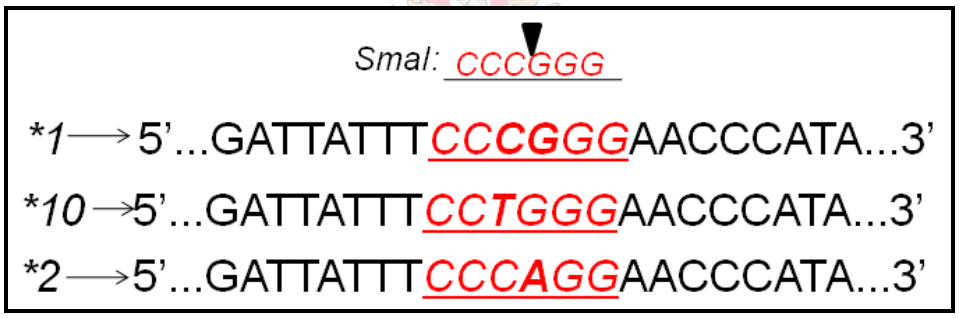


Figure 3.1: The removal of the *CCCGGG SmaI* recognition site as a result of both *CYP2C19*10* and *CYP2C19*2.*

With regards to the *in silico* analyses performed, the effect of non-synonymous mutations on the protein product was analysed using the sorting intolerant from tolerant (SIFT) (Ng and Henikoff, 2003), Polymorphism Phenotyping (PolyPhen) (Ramensky *et al.* 2002) and SNPs3D (http://www.snps3d.org/) algorithms, to analyse the impact of the non-synonymous variants on protein structure and function, as well as to determine the conservation across species. To elucidate the effect that the detected variants had on splicing, each of the exonic sequences and their surrounding intronic sequences were loaded into NetGene2 (Brunak *et al.* 1991; Hebsgaard *et al.* 1996) and SplicePort (Dogan *et al.* 2007) using the default parameters. The "wild type" or *CYP2C19*1* sequences were compared to the sequences containing the detected variants. Similarly, the "wild type" or *CYP2C19*1* 5'-upstream regions, with and without the detected variants, were

loaded into the PATCH, (vertebrates, min length = 6bp, core match>90%) (http://www.gene-regulation.com/cgi-bin/pub/programs/patch/bin/patch.cgi) MATCH (vertebrates, liver-specific, core match>0.9, matrix match>0.85) (Kel *et al.* 2003) and MatInspector (*Homo sapiens*, General Core Promoter Elements, vertebrates, core similarity>0.75, matrix similarity>0.75) (Quandt *et al.* 1995) software programs to determine the effect of these 5'-upstream variants on transcription factor binding sites. For the variants occurring in regions which have been deleted without significant effect on *in vitro* expression (Arefayene *et al.* 2003), only those variants creating additional transcription factor binding sites were regarded as significant. Furthermore significant results were required from all three programs. With regards to analysis of the effect of variants on mRNA secondary structure, mFold v3.2 (Zuker 2003) analysis was utilised on all transcribed areas at the set parameters. Finally, to calculate maximum-likelihood haplotypes from the observed data in the 15 sequenced individuals, Haploview 3.31 (Barret *et al.* 2005) was used. The tagger application was utilised at default settings to identify variants that were represented by other tagSNPs and did not require genotyping in the entire cohort.

3.3.7 Restriction Fragment Length Polymorphism (RFLP) Analysis

The frequencies of the prioritised variants detected were determined through the genotyping of an additional 35 or 85 healthy Xhosa DNA samples with the use of RFLP analysis. Potential restriction enzymes required for RFLP analysis were identified through the consultation of the *in silico* simulation of molecular biology experiments (Bikandi *et al.* 2004), RestrictionMapper (http://www.restrictionmapper.org) and WebCutter 2.0 (http://rna.lundberg.gu.se/cutter2/) websites to identify potential enzymes that would allow for the clear differentiation of all three possible genotypes. For all amplicons lacking suitable restriction enzyme sites or requiring costly or inefficient enzymes, mutagenic primers were designed to introduce restriction enzyme sites. Amplicons requiring mutagenic primers were amplified with the use of nested PCR reactions to avoid co-amplification of isoforms. All PCR reactions were performed as described in 3.2.2. For digestion reactions, verification with a positive control was performed whenever possible. In cases where no internal restriction enzyme recognition site was available, an additional sample, cut by the same enzyme, was digested in parallel to confirm the functionality of the enzyme. The restriction enzyme digest was performed by preparing a 20 µl reaction mix, consisting of a final concentration of 1X buffer, 2 U of enzyme, 5 µl PCR product and in selected digests, 1X BSA. The mixes were subsequently incubated in a water bath at the temperature and incubation time recommended by the manufacturer (refer to Table 3.3). The fragments obtained from the RFLP analyses were subsequently size fractionated and visualized with the use of gel electrophoresis on either 2-3% (w/v) ethidium bromide-stained agarose gels or 15% (w/v) non-denaturing polyacrylamide gel electrophoresis (PAGE) (5% cross-link) gels, according to the size of the fragments (refer to Table 3.3

for specifications). In the case of the agarose gels, 10 µl of restriction enzyme digest was loaded with 10 µl cresol loading buffer, while the PAGE gels were loaded with a mix of 5 µl restriction enzyme digest product and 5 µl bromophenol blue loading buffer. In both cases electrophoresis was performed using 1X TBE gel electrophoresis buffer, at 80 V and 200 V for the agarose and PAGE gels, respectively. Once PAGE analysis was complete, the gels were stained in a staining solution (1X TBE, 0.5 µg/ml Ethidium Bromide) for 10 min. Both the agarose and PAGE gels were subsequently analysed under UV light at A_{260} nm. The fragment sizes were determined with the use of Hyperladder IV and V (Bioline, London, UK) as molecular weight markers, according to the size of the fragments generated.

3.3.8 TaqMan® Copy Number Assays

In order to determine the number of *CYP2C19* copies present, a duplex real-time PCR was performed with a TaqMan® Copy Number Assay (Applied Biosystems™, Foster City, California, USA). The signal released by the *CYP2C19* probe was compared to the signal released by the reference probe. The TaqMan® Copy Number Assays were prepared to a final volume of 10 µl, consisting of 10 ng gDNA, a final concentration of 1X TaqMan® Genotyping Master Mix, 1X pre-designed TaqMan Copy Number Assay (Assay ID number: Hs02932336_cn) located in intron 6 of *CYP2C19* and 1X TaqMan Copy Number Reference Assay (*RNase PH1*) in a 384 well plate, using an EpMotion pipetting robot (Eppendorf, Hamburg, Germany), with four replicates for each sample. Thereafter the plate was sealed with optical adhesive film and loaded onto a 7900HT Fast Real-Time PCR system (Applied Biosystems™, Foster City, California, USA) at a reaction cycle of 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min. The target (*CYP2C19*) and reference (*RNase PH1*) gene amplicons present in the duplex real-time PCR were detected with FAM™ dye-labelled MGB and VIC® dye-labelled TAMRA™ probes, respectively (refer to Appendix 5). The results were then analysed on the CopyCaller™ Software (Applied Biosystems™, Foster City, California, USA), with the Most Frequent Sample Copy Number set as two and the manual cycle threshold (C_T) set to 0.2. A comparative C_T ($\Delta\Delta C_T$) relative quantitative analysis was performed by measuring the difference in C_T between the *CYP2C19* and *RNase PH1* genes and then comparing the ΔC_T values of all the samples.

Table 3.3: RFLP specifications

Enzymes, buffers and additives were supplied by New England Biolabs Inc., Beverly, USA

Variant	*CYP* Allele	Restriction Enzyme	Temperature (°C) and Additives	Genotype	Size of Fragments (bp)	Gel	Primer set
-2030C>T	Novel	*Mbo*I	37	CC CT TT	481, 20 501, 481, 20 501	2% Agarose	CYP2C19P+E1:mF2020 CYP2C19P+E1:R1
-2020C>A	Novel	*Dde*I	37	CC AC AA	328, 76, 66, 57, 16 404, 328, 76, 66, 57, 16 404, 66, 57, 16	2% Agarose	CYP2C19P+E1:F CYP2C19P+E1:R1
rs17882201		*Bse*NI	65	TT CT CC	367, 309, 198, 179, 79, 9 367, 309, 198, 179, 124, 79, 74, 9 367, 309, 179, 124, 79, 74, 9	2% Agarose	CYP2C19P+E1:F2 CYP2C19P+E1:R
IVS1-227	Novel	*Mse*I	37, BSA	AA AG GG	325, 208, 139, 55, 44 325, 252, 208, 139, 55, 44 325, 252, 139, 55	2% Agarose	CYP2C19E2+3:F CYP2C19E2+3:RS1
rs17880188		*HpyCH4*III	37	GG CG CC	293, 89, 41, 29 293, 89, 70, 41, 29 293, 89, 70	15% PAA	CYP2C19E8:F CYP2C19E8:R
rs12248569	**17*	*HpyCH4*V	37	CC AC AA	467, 40, 19 486, 467, 40, 19 486, 40	3% Agarose	CYP2C19P+E1:mF*17 CYP2C19P+E1:RS1
rs11568729		*Hha*I	37, BSA	CC CT TT	482, 44 526, 482, 44 526	3% Agarose	CYP2C19P+E1:F3 CYP2C19P+E1:RS1
rs17882687	**15*	*Bsa*BI	60	AA AC CC	264, 22 286, 264, 22 286	3% Agarose	CYP2C19P+E1:mF*15 CYP2C19P+E1:R
rs12769205		*Bfa*I	37	AA AG GG	222, 84, 72 306, 222, 84, 72 306, 72	2% Agarose	CYP2C19E2+3:FS1 CYP2C19E2+3:R
rs17884712	**9*	*HpyCH4*III	37	GG AG AA	447, 418, 98, 22, 11 545, 447, 418, 98, 22, 11 545, 418, 22, 11	3% Agarose	CYP2C19E2+3:F CYP2C19E2+3:R
rs4244285	**2*	*Sma*I	25	GG AG AA	410, 113 523, 410, 113 523	2% Agarose	CYP2C19E5:F CYP2C19E5:R
V347I	novel	*HpyCH4*IV	37	GG AG AA	305, 134, 123 305, 257, 134, 123 305, 257	2% Agarose	CYP2C19E7:F CYP2C19E7:R
rs4451645		*Nla*III	37, BSA	AA AG GG	204, 72, 6 204, 72, 54, 18, 6 204, 54, 18, 6	15% PAA	CYP2C19E9:mF CYP2C19E9:R
rs4986893	**3*	*Bsa*JI	60	GG AG AA	244, 124, 104 244, 228, 124, 104 244, 228	3% Agarose	CYP2C19E4:F CYP2C19E4:R
rs28399504	**4*	*HpyCH4*III	37	AA AG GG	339 339, 316, 23 316, 23	3% Agarose	CYP2C19P+E1:mF*4 CYP2C19P+E1:R
	**7*	*Bsa*AI	37	TT AT TT	271, 22 293, 271, 22 293	3% Agarose	CYP2C19E5:mF*7 CYP2C19E5:R
rs6413438	**10*	*Bsl*I	55	CC CT TT	108, 24, 10 132, 108, 24, 10 132, 10	15% PAA	CYP2C19E5:F CYP2C19E5:mR*10

3.3.9 Predicted Phenotype Classification

After genotyping the 100 Xhosa samples, classification according to predicted phenotype was performed. These classification groupings were PMs (two non-functional genes), IMs (one non-functional/two decreased function genes), EMs (two functional genes) or UMs (increased function genes, unaccompanied by non-functional/decreased function genes) (McKinnon and Evans 2000, Dandara *et al.* 2001; Ingelman-Sundberg *et al.* 2007).

3.3.10 Statistical Analysis

The genotype distribution for all identified SNPs was tested for Hardy-Weinberg equilibrium (HWE) using an analogue to Fisher's exact test in Tools For Population Genetic Analysis (TFPGA) Software v1.3 (Miller 1997). This was used in place of a Pearson chi-square (χ^2) analysis as all variants observed exhibited less than five individuals in at least one of the three genotype groups. Additionally, maximum-likelihood haplotypes were calculated from the observed data with the use of Haploview 3.31 (Barret *et al.* 2005).

3.4 Functional Analysis

3.4.1 Sequencing of Intervening 5'-upstream Unsequenced Area

Before dual reporter luciferase assays could be performed, the region of the promoter between the two previously sequenced areas (refer to Section 3.2.1 Figure 3.2), required sequence analysis to detect variants occurring within this region. The primers CYP2C19P+E1:F2 and CYP2C19P+E1:R2 (refer to Table 3.1) were used to amplify the desired region from the P+E1 template DNA and semi-automated sequence analysis was performed in the same manner as described in 3.2.4. The variants detected in this region were analysed with PATCH, MATCH and MatInspector software programs, as specified in 3.2.6.2, to identify putative transcription factor binding sites which were created as a result of these variants. Only the creation of transcription factor binding sites was considered due to the fact that it has been reported that the deletion of this area has no impact on the expression of CYP2C19 (Arefayene *et al.* 2003). The detected variants were subsequently genotyped in 50 individuals to determine their LD to previously detected variants in the population. The genotyping was performed using mutagenic primers (refer to Table 3.4) in combination with RFLP analysis (refer to Table 3.5).

Figure 3.2: Region requiring sequence analysis for dual luciferase reporter assays.
The region which requires sequence analysis is depicted as "unsequenced area" in this Figure.

Table 3.4: Primer sequences for genotyping of additional 5' variants

Region	Template Primers (5' – 3')	
P+E1	CYP2C19P+E1:F	CAG AAC TGG AAC ACC TAG CTC TCA
P+E1	CYP2C19P+E1:R	GAC AGA CTG GAA AAG GCA ACA AAA G'
Region	**Internal Primers (5' – 3')**	
P+E1	CYP2C19P-1439mF	CTT AAT AAG AGA ACT GGA AAT AAC C**gt** A
P+E1	CYP2C19P-1418mF	CCT CAT TAG GAA ATT TAG AAC AA**g** TA
P+E1	CYP2C19P-1041mF	GCT CTT CCT TCA GTT ACA CTG A**a**C
P+E1	CYP2C19PRS	GAG ATG CTT TGA GAA CAG AAG ACA C

P: promoter, E: exon, F: forward primer, R: reverse primer, m: mutagenic primer, lower case letter: mutagenic bases

Table 3.5: RFLP specification for additional 5' variants

Variant	Restriction Enzyme	Temperature (°C)	Genotype	Size of Fragments (bp)	Gel	Primer set
rs17878739	*Rsa*I	37	TT CT CC	667 667, 640, 27 640, 27	2% Agarose	CYP2C19P-1439mF CYP2C19PRS
rs3814637	*Rsa*I	37	CC CT TT	619, 25 644, 619, 25 644	2% Agarose	CYP2C19P-1418mF CYP2C19PRS
rs7902257	*Acl*I	37	AA AG GG	265 265, 242, 23 242, 23	3% Agarose	CYP2C19P-1041mF CYP2C19PRS

3.4.2 Primer Design and Amplification

For ligation into the pGL4.10 [luc2] vector (Promega, Madison, Wisconsin, USA) (refer to Appendix 6 for vector map), primers were designed with 5' overhangs incorporating *Xho*I and *Bgl*II restriction endonuclease recognition sites into the forward and reverse primers respectively, with an additional five nucleotides to allow for efficient enzyme activity (refer to Table 3.6). These primers were used to amplify three characterised samples in order to create the desired fragments for insertion into the pGL4.10 vectors (depicted in Figure 3.3), however, as these primers were not specific for *CYP2C19*, an initial amplification reaction was performed using the previously mentioned CYP2C19P+E1:F and CYP2C19P+E1:R primers. All fragments amplified for insertion into the vectors were designed to include the 5'-upstream region ranging from the nucleotide directly upstream from the ATG initiation codon, to 2 095 bp upstream from the ATG codon. This eliminated the need for a minimal promoter and ensured that the region containing the -2030C>T and -2020C>A was included in these fragments.

Table 3.6: Primer sequences for luciferase constructs

Primers	
CYP2C19pGL4.10:F	5'- CCC CC**C TCG AG**[1]C AGA ACT GGA ACA CCT AGC TCT C -3'
CYP2C19pGL4.10:R	5'- CCC CC**A GAT CT**[2]T GAA GCC TTC TCC TCT TGT TAA GAC AAC C -3'
RV3 (Promega)	5'-CTA GCA AAA TAG GCT GTC C-3'

[1]Underlined sequence indicates incorporated *Xho*I restriction site
[2]Underlined sequence indicates incorporated *Bgl*II restriction site

To minimise the occurrence of errors during the amplification process, DNA Taq Elongase® Enzyme (Invitrogen™, Carlsbad, California, USA), containing proofreading activity, was utilised. For the initial amplification, a reaction mix was prepared to a final volume of 25 µl, containing 15 ng of gDNA, a final concentration of 1X Invitrogen buffer B, 0.4 mM dNTPs, 0.2 µM of each primer and 0.5 U Taq. The amplification cycle reaction was performed at an initial denaturation of 94°C for 30 sec; followed by 35 cycles of denaturation at 94°C for 30 sec and an annealing and extension step at 68°C for 3 min 30 sec; concluding with a final extension step at 68°C for 5 min. Nested PCR reactions were subsequently performed using 1/100 dilutions of the DNA template. The reaction mix for the nested PCRs utilised the same reagents and concentrations as the initial PCR, at a final concentration of 50 µl, while the amplification cycle reaction was performed at an initial denaturation of 94°C for 30 sec; followed by 35 cycles of denaturation at 94°C for 30 sec, annealing at 55°C for 30 sec and extension step at 68°C for 2 min; concluding with a final extension step at 68°C for 5 min.

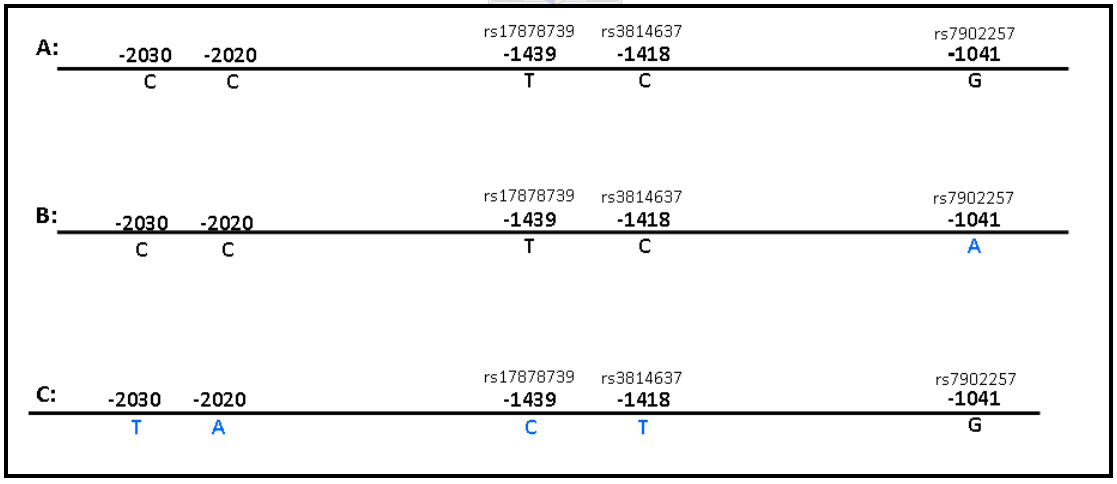


Figure 3.3: Fragments inserted into pGL4.10 vectors.

Fragment A contains only the ancestral nucleotides as described by NCBI (NM 000769), fragment B contains the sequence as described by Ensembl (ENSG00000165841) and fragment C contains the variants of interest as well as any variants occurring on the same allele as these variants. Variant nucleotides are depicted in blue.

3.4.3 Preparation and Ligation of the Constructs and Vectors

Following the successful amplification of the desired fragments, a purification reaction was performed with MSB® Spin PCRapace columns (Invitek Inc. GmbH, Berlin, Germany) (refer to Appendix 3.4), after which the purified products and pGL4.10 vectors were digested with both the *XhoI* and *BglII* restriction enzymes (Fermentas, Glen Burnie, Maryland, USA). The digestion reactions contained 10 U of each enzyme and a final concentration of 2X buffer 2 (Fermentas, Glen Burnie, Maryland, USA). The digestion reaction mix for the vector was prepared to a final volume of 20 µl, containing 4 µg of the pGL4.10 vector, while the reaction mix for the purified PCR products was prepared to a final volume of 50 µl, containing 25 µl of the PCR product. Once the digestion reaction was complete, the resulting products underwent gel electrophoresis for 1 hr 30 min at 80 V on a 0.8% (w/v) agarose gel, visualisation under a UV light at A_{260} nm and excision of the desired fragments from the gel, using a sterile blade. The excised fragments were subsequently purified with a QIAquick gel extraction kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol (refer to Appendix 3.5) and the concentrations of the purified samples were determined with a spectrometer (NanoDrop® ND-100, Nanodrop Technologies Inc., Wilmington, Delaware, USA) by measuring the absorbance of the amplified DNA at 260 nm. For the ligation reactions, a 1:3 vector:insert ratio was achieved in a reaction mix of 30 µl, containing 100 ng of vector, 1 U of T4 DNA ligase and a final concentration of 1X ligase buffer (Invitrogen™, Carlsbad, California, USA). The ligation reactions were subsequently incubated at 4°C overnight.

3.4.4 Transformation Reactions and Sequence Confirmation

The resulting ligation reactions were transformed into *E.cloni*® Chemically Competent cells (Lucigen Corporation, Middleton, Wisconsin, USA) according to the manufacturer's protocol (refer to Appendix 3.6). Once transformation was complete, 600 µl of recovery medium was added to each tube and left to shake at 250 rpm for 1 hr at 37°C. Subsequently, 100 µl of the resulting transformed cells were plated onto LB agar plates (Lucigen Corporation, Middleton, Wisconsin, USA) containing 50 µg/ml amplicon, prepared according to manufacturer's instructions (refer to Appendix 3.7) and plates were incubated overnight at 37°C. After the successful transformation and formation of the respective colonies, 20 white recombinant colonies from each representative agar plate were picked with a sterile toothpick, re-plated and grown on a master plate. Ten colonies from each resulting master plate were subsequently picked once again with a sterile toothpick and genotyped for either -2030C>T for construct C or rs7902257 for construct B, using the PCR-RFLP conditions previously described in Section 3.2.7. Once representatives for each desired construct were identified, the orientation of the inserted fragments was verified through amplification with the vector specific RV3 forward primer in combination with the insert specific CYP2C19pGL4.10:R reverse primer (refer to Table 3.4 for primer sequences). The resulting PCR products were subjected to gel electrophoresis

on a 1% (w/v) agarose gel, after which successfully amplified products were prepared for semi-automated sequence analysis. After sequence confirmation, the colonies containing the constructs of interest were inoculated in LB medium containing 50 µg/ml ampicilin, left to shake at 37°C, 250 rpm, overnight and subsequently extracted using a GenElute™ Plasmid Miniprep Kit (Sigma-Aldrich (Pty) Ltd, Aston Manor, South Africa) according to the manufacturer's protocol (refer to Appendix 3.8).

3.4.5 Cell Culture

Human hepatocarcinoma liver cell line HepG2 cells from the American Type Culture Collection (ATCC No. HB-8065, Rockville, Minnesota, USA) were cultured in three independent 75 cm^2 sterile culture flasks (Cellstar®, Greiner Bio-One, Frickenhausen, Germany) to allow for the dual luciferase assay experiments to be performed in triplicate. The cells were allowed to grow at 37°C, 5% CO_2, in an Heraeus Cell 150 incubator (Kendro Laboratory Products, Asheville, North Carolina, USA) in Dulbecco's Modified Eagle's Medium (DMEM) with 4 500 mg/l glucose, 110 mg/l sodium pyruvate and L-glutamine, supplemented with 10% fetal bovine serum (FBS) (v/v), 1X MEM non-essential amino acid solution and Pen Strep. The medium was replenished as required. Once the cells reached a confluency of approximately 80%, the medium was removed, the cells were rinsed with Hanks Balanced Salt solution and resuspended in 5 ml Trypsin EDTA, after which they were sub-cultured. All reagents were supplied by Sigma-Aldrich (Pty) Ltd.

3.4.6 Transfection and Passive Lysis Reactions

For the transfection reactions, sufficient wells were prepared to allow for each experiment to be performed in triplicate, with three replicates for each experiment. Therefore, 1 x 10^5 HepG2 cells per well were seeded into enough wells to allow for this. Each experimental plate was prepared separately using an independent flask of HepG2 cells, after which the plates were placed in an Heraeus Cell 150 incubator (Kendro Laboratory Products, Asheville, North Carolina, USA) at 37°C, 5% CO_2, for 24 hr. In the meantime 210 ng of DNA for each well, consisting of 200 ng construct and 10 ng pGL4.73 vector [hRluc/SV40] which was used as a control for transfection efficiency (refer to Appendix 6 for vector map). A 1:50 ratio of GeneJuice® Transfection Reagent:DMEM (Sigma-Aldrich (Pty) Ltd, Aston Manor, South Africa), with the GeneJuice® amounting to 1.1 times the DNA concentration was subsequently prepared, vortexed and incubated at room temperature for 5 min, after which it was added to the DNA mix, incubated at room temperature for 15 min and added to the wells containing the HepG2 cells. The plate was subsequently returned to the incubator at 37°C, 5% CO_2, for 24 hr, after which the old medium was discarded and the cells were rinsed with 500 µl PBS to facilitate the detachment of dead cells. Thereafter, the PBS was discarded and replaced with

100 µl of lysis buffer. After 15 min of gentle shaking at room temperature, the plates were stored at -80°C overnight.

3.4.7 Dual Reporter Luciferase Assays

In order to determine the activity of the promoter area of interest, the dual luciferase assay (Promega, Madison, Wisconsin, USA) measures the firefly luciferase activity of cells transfected with reporter pGL4.10 vectors carrying the *luc2* gene as well as the renilla luciferase activity of the same cells co-transfected with the pGL4.73 vector carrying the *hRluc* gene and SV40 promoter. Luminescence was measured on a GloMax™ 96 Plate Luminometer (Promega, Madison, Wisconsin, USA), using LARII reagent to initiate the firefly luciferase activity and Stop&Glo reagent to quench the firefly luciferase activity and initiate the renilla firefly activity. The values were measured at 10 sec intervals with no pre-measurement delay. Values obtained for the firefly luciferase activity were subsequently normalised through division by the values obtained for the renilla luciferase activity, in the same culture dish.

3.4.8 Statistical Analyses

After the data were obtained, the mean and standard deviation for each construct was calculated and outliers were removed. The values obtained for each construct were then tested for normality using a Shapiro-Wilk test (Shapiro and Wilk 1965), while equal variances were tested using an F-test. Significant differences were determined using a two–sample *t*-test. Differences in fold induction ±SEM were regarded as significant at $p<0.01$

3.5 *In Silico* Analysis of the 5' Region

3.5.1 Comparative Sequence Analysis of the 5'-upstream Region

The entire 5'-upstream regions of all four *Homo sapiens CYP2C* genes, as well as 5'-upstream regions from *CYP2C19* orthologues, were obtained through consultation with the NCBI and Ensembl websites (refer to Table 3.7). These regions of interest ranged from the last stop codon of the previous gene to the first initiation codon of the *CYP2C19* orthogue/paralogue of interest. Before the extracted sequences could be aligned to the *CYP2C19* 5'-upstream region, they were separated into suitable sizes and groups to fit the rVISTA platform (Loots *et al.* 2002) requirements. The groups of sequences were then loaded onto rVista and the repeat masker for human/primates was activiated to determine whether the regions of similarity occurred within repetitive elements. Once the data was retrieved, the sequences were analysed to identify regions showing more than 70% sequence similarity over more than 100 bp to the 5'-upstream region of *CYP2C19*. Using these areas of high sequence similarity, consensus sequences were created by ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) and all extracted 5'-upstream regions were re-aligned to the consensus

sequences created by ClustalW2, to identify regions of similarity missed by the original sequence comparisons. The data from these analyses were subsequently prioritised by identifying regions which aligned to more than two species, excluding the chimpanzee. Finally, these areas were analysed to identify potential transcription factor binding sites occurring within the prioritised regions with the use of PATCH, MATCH and MatInspector, with the parameters specified in 3.2.6.2.

Table 3.7: 5' regions of genes used for comparative sequence analysis

Accession number	Downstream Gene	Length (bp)	Species	Common Name
NM_000769	*CYP2C19*	27 267	*Homo sapiens*	Human
NM_000771	*CYP2C9*	85 774	*Homo sapiens*	Human
NM_000770	*CYP2C8*	124 904	*Homo sapiens*	Human
NM_000772	*CYP2C18*	35 837	*Homo sapiens*	Human
NW_001471715.1	*CYP2H1*	37 213	*Gallus gallus*	Chicken
NW_001220741.1	*CYP2C19*	30 711	*Pan troglodytes*	Chimpanzee
NW_001494355.2	*CYP2C87*	19 025	*Bos Taurus*	Cow
NW_001581859.1	*CYP2C75*	30 112	*Monodelphis domestica*	Opossum
NW_001877590.1		30 003	*Danio rerio*	Zebra fish
NW_876285.1	*CYP2C21*	20 437	*Canis famililiaris*	Dog
NC_004354.3	*CYP18a1*	17 260	*Drosophila melanogaster*	Fruit fly
NW_001867386.1	*CYP2C92*	55 221	*Equus caballus*	Horse
NT_039687.7	*CYP2C66*	30 015	*Mus musculus*	Mouse
NW_001885525.1	*CYP2C49*	7 397	*Sus scrota*	Pig
NW_047565.2	*CYP2C24*	27 260	*Rattus norvegicus*	Rat
NW_001124218.1	*CYP2C75*	46 619	*Macaca mulatta*	Rhesus monkey

3.5.2 CpG Island Analysis

All four *CYP* genes were analysed for the presence of CpG islands in the region extending from the last translational trinucleotide of the 5'-upstream gene to the first translated trinucleotide of the 3'-downstream gene. The analysis was performed by consulting the CpG island searcher (http://www.cpgislands.com/) and CpG Plot (http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html) software programs. Analysis with CpG island searcher used the parameters suggested by Takai and Jones (2002) (%GC = 55%, ObsCpG/ExpCpG = 0.65, length = 500 bp, gap between adjacent islands = 100 bp), while CpGPlot analysis used the set parameters, excluding the minimal length which was changed from 200 bp to 500 bp. Analysis was performed to determine if any correlation could be made between the differences in expression observed between the four genes and the number or position of CpG islands observed within and around the genes.

CHAPTER 4:

RESULTS

CHAPTER 4: RESULTS

4.1 Identification of Variants Occurring in the Xhosa Population

4.1.1 Variant Detection

Successful PCR amplification and bi-directional re-sequencing of the *CYP2C19* amplicons, encompassing the 5'-upstream regions, exonic regions, exon-intron boundaries as well as the 3'UTR, in 15 healthy Xhosa individuals revealed 27 variants. The details of these variants are summarised in Table 4.1, while sequence chromatograms are shown in Appendix 7. Although the initial amplification of PCR fragments larger than 1 kb in size proved problematic, this was rectified by performing new DNA extractions for all samples of inferior gDNA quality. Additionally, some problems were initially encountered with regards to the co-amplification of *CYP2C9* due to the high level of sequence similarity observed between the genes, but this was solved either through the use of nested PCR reactions, or through the design of primers with as many nucleotide differences to *CYP2C9* as possible, in combination with high annealing temperatures utilised in the PCR cycles. The *CYP2C19* specific amplification of all amplification products was verified through the sequence analysis of 15 individuals for each amplicon under inspection.

With regards to the previously characterised *CYP2C19* alleles (http://www.cypalleles.ki.se/cyp2c19.htm), variants from four different *CYP2C19** alleles were detected in this population, namely *CYP2C19*2C*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17*. Although not all the variants described by the Human *CYP2C19* Allele Nomenclature website as forming part of the *CYP2C19*2C* allele were detected, six of the *CYP2C19*2C* variants occurring in LD with the allele-defining *CYP2C19*2* variant, were indeed detected in this population. The allele-defining variants described on the Human *CYP2C19* Allele Nomenclature website for the *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17* alleles were all detected within the Xhosa population. In the case of *CYP2C19*9* and *CYP2C19*15*, additional variants occurring in perfect LD with the allele-defining variants were identified, which have not previously been described on the Human *CYP2C19* Allele Nomenclature website.

In addition to the documented *CYP2C19* allelic variants detected in this population, five novel variants were detected. Two of these novel variants were detected in intronic regions, namely intron 1 and 2; however, *in silico* analysis (discussed in more detail in Section 4.1.2) did not predict any likely functional effect on CYP2C19 as a result of these variants. Of greater value to CYP2C19 functioning, were the remaining novel variants detected. One of these variants was found in exon 7 and was shown to result in an amino acid change from valine to isoleucine at position 374, thereby affecting the protein code of CYP2C19 and warranting further inspection. This variant was designated *CYP2C19*28* by the Human *CYP2C19* Allele Nomenclature website. The last two novel variants were

both found in the 5'-upstream region of *CYP2C19* and were predicted to remove a transcription factor binding site involved in the expression of *CYP2C19* (discussed further in Section 4.1.2.).

Although rs numbers have been allocated to the remaining variants, they have not been described on the Human *CYP2C19* Allele Nomenclature website. None of the *in silico* analyses performed (described in Section 4.1.2) predicted a functional affect as a result of these variants.

Table 4.1: The variants detected in the Xhosa cohort

Allele	Variant	Region	Effect on protein	Effect on enzyme activity	Allele frequency (n = 15)	Genotype frequency (n = 15)		Allele frequency (n = 50/100)	Genotype frequency (n = 50/100)	
*CYP2C19*2*	**- 98T>C rs4986894**	5'			0.167	TT	0.67			
						TC	0.33			
						CC	0.00			
	IVS1-231G>A rs7916649	Intron 1			0.767	GG	0.07			
						GA	0.33			
						AA	0.60			
	681G>A rs4244285	Exon 5	Splicing defect	Defective	0.167	GG	0.67		GG	0.62
						GA	0.33	0.21 (100)	GA	0.35
						AA	0.00		AA	0.03
	IVS5+228G>A rs12571421	Intron 5			0.167	GG	0.67			
						GA	0.33			
						AA	0.00			
	IVS5-51C>G rs4417205	Intron 5			0.200	CC	0.67			
						CG	0.27			
						GG	0.07			
	IVS6-196T>A rs28399513	Intron 6			0.167	TT	0.67			
						TA	0.33			
						AA	0.00			
	990C>T rs3758580	Exon 7	V > V		0.133	CC	0.67			
						CT	0.27			
						TT	0.07			
*CYP2C19*9*	**431G>A rs17884712**	Exon 3	R > H	Decrease *in vitro*	0.167	GG	0.73		GG	0.84
						GA	0.20	0.09 (100)	GA	0.15
						AA	0.07		AA	0.01
	IVS1+11T>C rs17882201	Intron 1			0.167	TT	0.73		TT	0.88
						TC	0.20	0.07 (50)	TC	0.10
						CC	0.07		CC	0.02
*CYP2C19*15*	**- 2030C>T**	5'			0.067	CC	0.87		CC	0.78
						CT	0.13	0.11 (50)	CT	0.22
						TT	0.00		TT	0.00
	- 2020C>A	5'			0.067	CC	0.87		CC	0.78
						CA	0.13	0.11 (50)	CA	0.22
						AA	0.00		AA	0.00
	-1439T>C rs17878739	5'			0.067	TT	0.87		TT	0.78
						TC	0.33	0.11 (50)	TC	0.22
						CC	0.00		CC	0.00
	55A>C rs17882687	Exon 1	I > L		0.067	AA	0.87		AA	0.82
						AC	0.13	0.09 (100)	AC	0.18
						CC	0.00		CC	0.00
*CYP2C19*17*	**-806C>T rs12248560**	5'		Increase	0.033	CC	0.93		CC	0.81
						CT	0.07	0.10 (100)	CT	0.18
						TT	0.00		TT	0.01
*CYP2C19*28*	**1120G>A**	Exon 7	V > I		0.033	GG	0.93		GG	0.99
						GA	0.07	0.01 (100)	GA	0.01
						AA	0.00		AA	0.00

Allele	Variant	Region	Effect on protein	Effect on Enzyme activity	Allele frequency (n = 15)	Genotype frequency (n = 15)		Allele frequency (n = 50/100)	Genotype frequency (n = 50/100)	
	-1418C>T rs3814637	5'			0.167	CC	0.67	0.22 (50)	CC	0.64
						CT	0.33		CT	0.28
						TT	0.00		TT	0.08
*CYP2C19*27*	**-1041G>A rs7902257**	5'		Decrease in expression *in vitro*	0.167	GG	0.67	0.33 (100)	GG	0.64
						GA	0.33		GA	0.28
						AA	0.00		AA	0.08
	-783C>T rs11568729	5'			0.067	CC	0.87	0.08 (100)	CC	0.85
						CT	0.13		CT	0.15
						TT	0.00		TT	0.00
	99C>T rs17885098	Exon 1	None	None	0.833	CC	0.00			
						CT	0.33			
						TT	0.67			
	IVS1+20C>T rs17881883	Intron 1			0.033	CC	0.93			
						CT	0.07			
						TT	0.00			
	IVS1-340T>C rs17884832	Intron 1			0.100	TT	0.80			
						TC	0.20			
						CC	0.00			
Novel	**IVS1-227G>A**	Intron 1			0.033	GG	0.93	0.05 (50)	GG	0.90
						GA	0.07		GA	0.10
						AA	0.00		AA	0.00
Novel	**IVS2-48C>T**	Intron 2			0.033	CC	0.93			
						CT	0.07			
						TT	0.00			
	IVS2-23A>G rs12769205	Intron 2			0.200	AA	0.67	0.22 (100)	AA	0.60
						AG	0.27		AG	0.36
						GG	0.07		GG	0.04
	IVS4-40T>C rs57752480	Intron 4			0.067	TT	0.87			
						TC	0.13			
						CC	0.00			
	IVS7-106T>C rs4917623	Intron 7			0.033	TT	0.93			
						TC	0.07			
						CC	0.00			
	IVS8+122G>C rs17880188	Intron 8			0.067	GG	0.87	0.03 (50)	GG	0.94
						GC	0.13		GC	0.06
						CC	0.00		CC	0.00
	IVS8-17A>G rs4451645	Intron 8			0.067	AA	0.87	0.03 (100)	AA	0.94
						AG	0.13		AG	0.06
						GG	0.00		GG	0.00
	IVS8-119C>T rs12268020	Intron 8			0.167	CC	0.67			
						CT	0.33			
						TT	0.00			

Numbers in brackets are the number of individuals genotyped, as described in Section 3.2.6.
Only those effects that have been reported on the Human *CYP2C19* Allele Nomenclature website have been recorded in this table.

4.1.2 Prioritisation of Detected Variants

Due to the large amount of variation detected in the *CYP2C19* gene in the Xhosa population, a prioritisation strategy was required to decrease the number of variants that required genotyping in a larger cohort of healthy Xhosa individuals. In order for the successful implementation of this prioritisation strategy, initial *in silico* analyses were utilised to determine the likely functionality of novel/uncharacterised SNPs. In addition to this, all non-synonymous and allele-defining variants were identified for inclusion in the genotyping platform due to their non-synonymous nature or functional evidence reported by the Human *CYP2C19* Allele Nomenclature home page. The variants that were prioritised in this manner were the novel V374I SNP (*CYP2C19*28*), as well as the allele-

defining *CYP2C19*2* (rs4244285), *CYP2C19*9* (rs17884712), *CYP2C19*15* (rs17882687) and *CYP2C19*17* (rs12248560) variants. The allele-defining *CYP2C19** variants were chosen to represent each of the *CYP2C19** alleles as previous characterisation of these alleles has demonstrated that the effect of the respective alleles is dependent on these variants. Due to the recognised format of the *CYP2C19** alleles, from this point on, the allele-defining variants shall be referred to using their *CYP2C19** allele names. Additionally, the novel V374I SNP was genotyped in the entire cohort by virtue of its non-synonymous properties. The *in silico* analysis performed on this variant to determine the putative extent of the effect exerted on the CYP2C19 protein, did not however predict any significant results. All three algorithms used; namely SIFT, PloyPhen and SNPs3D; gave tolerant (>0.201), benign (<0.999) and non-deleterious (positive support vector machine (svm)) scores, respectively.

In an attempt to identify variants potentially affecting the splicing of CYP2C19, *in silico* splice site analysis was performed for all detected variants occurring between the start and stop codons of *CYP2C19*. Although the analysis from both NetGene2 and SplicePort did confirm aberrant splicing as a result of the previously characterised *CYP2C19*2* and *CYP2C19*7* (rs12571421) SNPs, which have been proven to result in the creation of a new acceptor site (De Morais *et al.* 1994a) and the removal of a donor site (Ibeanu *et al.* 1999), these analyses did not reveal any other significant changes in splice sites as a result of the other variants detected in the Xhosa. Table 4.2 demonstrates that when the sequence for exon 5 and the surrounding intronic regions is loaded into both splice site prediction programs, the correct position for the acceptor site is predicted for *CYP2C19*1*, however, when the same sequence with the *CYP2C19*2* (19154G>A) mutation is loaded, a new stronger acceptor site, occurring in exon 5 instead of in the exon-intron boundary, is predicted. Conversely, when the same sequence is loaded containing the *CYP2C19*7* (19294T>A) mutation, the donor site is predicted to be removed as a result of the mutation.

Table 4.2: Variants affecting splice sites.

Program	Allele	Confidence	Intron-exon sequence (5'-3')	Effect
NetGene2	**1* (19154G)	0.24	TTTCTCTTAG^ATATGCAATA	
	**2* (19154A)	0.34	TATTTCCCAG^GAACCCATAA	Creates new acceptor site
SplicePort	**1* (19154G)	0.08	CTCTTAG^ATATG	
	**2* (19154A)	0.13	TTCCCAG^GAACC	Creates new acceptor site
NetGene2	**1* (19294T)	0.95	AATGGAGAAG^GTAAAATGTT	
	**7* (19294A)	0	TTTAATAAAG^GTATAGATAC	Removes donor site
SplicePort	**1* (19294T)	1.59	AGAAG^GTAAAAT	
	**7* (19294A)	no predicted sites		Removes donor site

**1*: "wild type allele", **2* and **7*: non-functional alleles
The specific mutations affecting the splice sites are given in brackets

To identify the potential creation/abolishment of any transcription factor binding sites occurring as a result of the variants detected in the 5'-upstream region of *CYP2C19*, *in silico* transcription factor binding site analysis was utilised. This analysis, convincingly predicted the removal of a GATA factor binding site as a result of the novel -2030C>T variant (Refer to Table 4.3). As the -2030C>T and -2020C>A variants were shown to occur in perfect LD with one another, transcription factor binding site analysis was performed for both SNPs in combination with each other, however it was only the -2030C>T variant which was predicted to affect the GATA transcription factor binding site.

The last form of *in silico* analysis performed for the prioritisation of the detected variants was the use of mFold analysis on all variants occurring within the transcribed regions of *CYP2C19*. Although this analysis did reveal a slight change in folding structure of *CYP2C19* as a result of the rs7916649, IVS1-227, IVS2-48, V374I, rs4917623 and rs17880188 variants, only rs28399513 showed any noteworthy change in folding, as shown in Figure 4.1. However, as rs28399513 occurs after the main *CYP2C19*2* (rs4244285) splicing defect, the effect of the aberrant folding is lost due to the truncated nature of this allele.

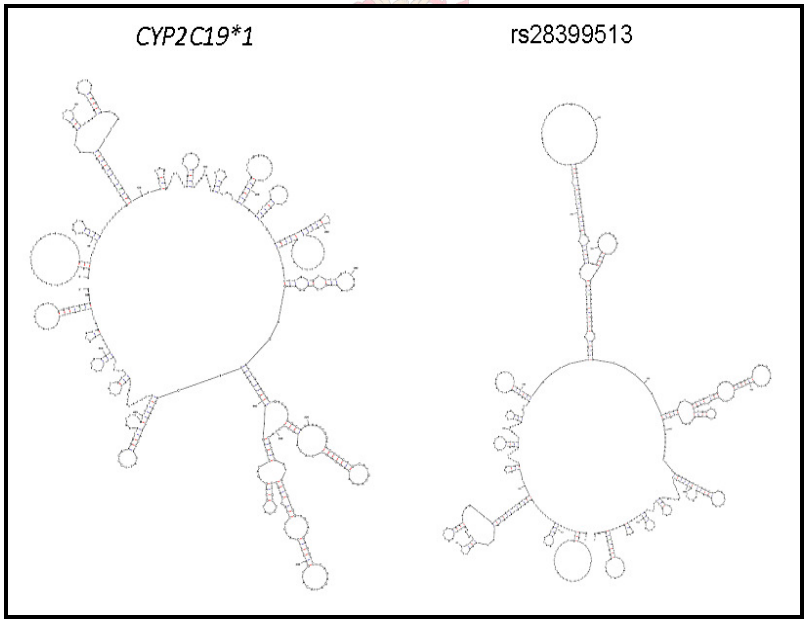


Figure 4.1: Predicted differences in mRNA folding observed between the *CYP2C19*1* allele and rs28399513.

Table 4.3: Effect of the -2030C>T variant on transcription factor binding sites

Program		Matrix identifier	Core match	Matrix match	Factor name	Sequence
MATCH	Removes	V$GATA_C	1.000	0.982	GATA-X	caagcTTAT**C**t
		V$GATA1_02	1.000	0.962	GATA-1	aagctTAT**C**Ttacc
		V$GATA1_03	1.000	0.941	GATA-1	aagcTTAT**C**ttacc
		V$GATA1_05	1.000	0.942	GATA-1	gcTTAT**C**tta
		V$GATA1_06	1.000	0.944	GATA-1	gcTTAT**C**tta
		V$GATA2_02	1.000	0.977	GATA-2	gcTTAT**C**tta
Program		**Identifier**	**Mismatches**	**Score**	**Binding Factor**	**Sequence (Search Pattern)**
PATCH	Removes	HS$GG_29	0.000	100	GATA-1	TTAT**C**T
		HS$AG_02	0.000	100	GATA-1	TTAT**C**T
		MOUSE$EPOR_01	0.000	100	GATA-1	TTAT**C**T
		HS$AG_14	0.000	100	GATA-1	TTAT**C**T
		HS$TCRBL_10	0.000	100	GATA-3	TTAT**C**T
		HS$BG_31	0.000	100	GATA-1	TTAT**C**T
		MOUSE$TCRBL_03	0.000	100	GATA-3	TTAT**C**T
		MOUSE$PBGD_06	0.000	100	GATA-1	A**G**ATAA[1]
		HS$AG_09	0.000	100	GATA-1	A**G**ATAA[1]
		MOUSE$MCCPA_01	0.000	100	GATA-1, GATA-2	A**G**ATAA[1]
		RAT$BNP_01	0.000	100	GATA-4	A**G**ATAA[1]
		RAT$BNP_02	0.000	100	GATA-4	A**G**ATAA[1]
		MOUSE$PDGFRA_01	0.000	100	GATA-4	A**G**ATAA[1]
		HS$ET1_01	0.000	100	GATA-2	TTAT**C**T
Program		**Matrix**	**Core similarity**	**Matrix similarity**	**Detailed Family Information**	**Sequence**
MatInspector	Removes	V$GATA1.03	1.000	0.974	GATA binding factors	gtaa**G**ATAagctt[1]

Nucleotides underlined and in bold are those nucleotides affected by the -2030C>T variant, [1] reverse complement sequences

After the *in silico* analyses were performed, all potentially functional variants were identified along with the non-synonymous variant and *CYP2C19** allele-defining variants. The data from the 15 sequenced Xhosa individuals were then analysed in Haploview 3.31 (Barret *et al.* 2005) to determine the extent of LD observed between the detected variants. Figure 4.2 shows how the *CYP2C19*2*, *CYP2C19*9* and *CYP2C19*15* alleles differ from those described by the Human *CYP2C19* Allele Nomenclature website. With regards to *CYP2C19*2,* not all the variants described as belonging to this allele were detected, however, in the case of *CYP2C19*9* and *CYP2C19*15* additional variants were detected occurring in perfect LD with the allele-defining variants, namely rs17882201 for *CYP2C19*9* and -2030C>T and -2020C>A for *CYP2C19*15*. These variants were genotyped in an additional 35 individuals to determine whether they remained in perfect LD with the allele-defining SNPs, which was found to be the case. Furthermore, the variants identified to potentially affect *CYP2C19* were forced in the tagger application of Haploview 3.31 to ensure that they were included in the subsequent genotyping strategy utilised. The tagger application identified which SNPs would be represented by the forced tagSNPs and did not require further genotyping. For a table displaying all variants that have been prioritised for genotyping in a larger cohort, refer to Table 4.4.

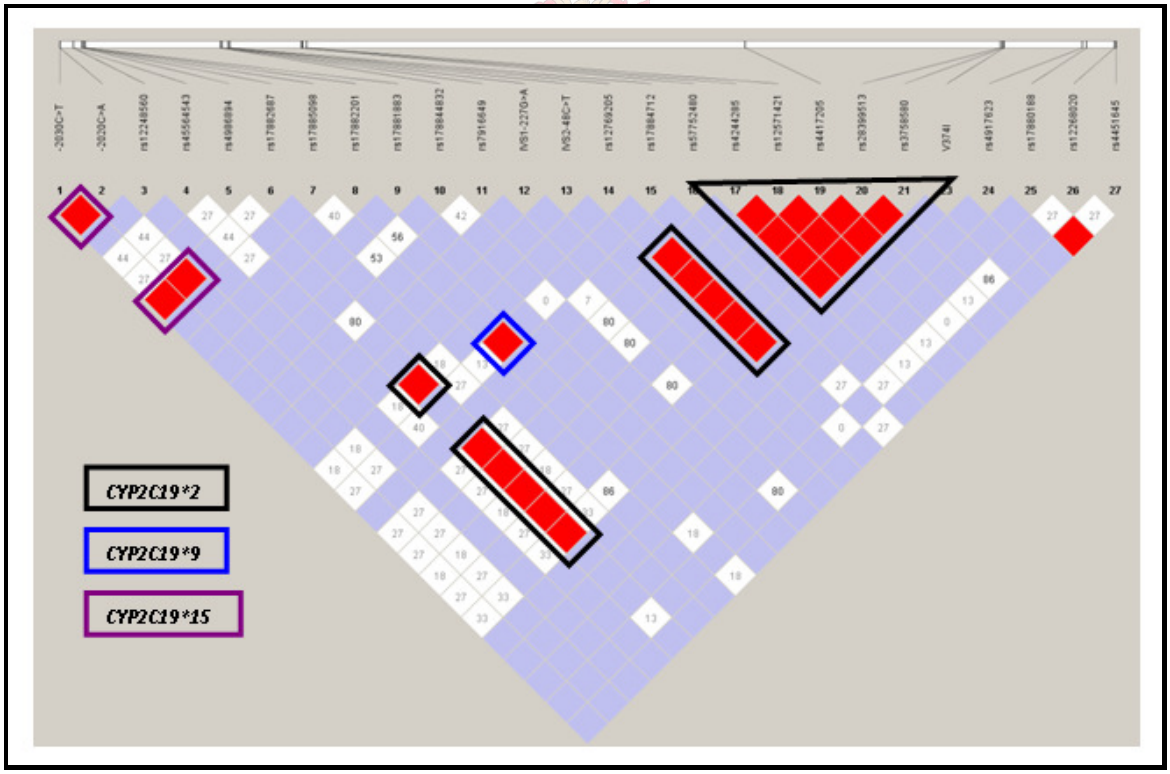


Figure 4.2: Haplotype analysis of the 15 sequenced Xhosa individuals.
The different *CYP2C19** alleles detected in the Xhosa population are depicted by the coloured blocks.

Table 4.4: Variants prioritised for genotyping in a larger cohort

SNPs prioritised for genotyping	**rs12248560(*17)[1]**	**rs17882687(*15)[1]**	**rs17884712(*9)[1]**	**rs4244285(*2)[1]**	**rs4451645[1]**	**V374I[1]**	**rs45564543[1]**
SNPs captured by prioritised SNPs using the tagger application		**-2030C>T[2]** **-2020C>A[2]**	**rs17882201[2]**	**rs12769205[2]** rs4417205 rs28399513 rs12571421 rs4986894	**rs17880188[2]**	**IVS1-227G>A[2]**	
Intronic SNPs	rs17884832	rs17881883	rs7916649	rs4917623	rs12268020	rs57752480	IVS2-48C>T
*1 SNPs	rs17885098	rs3758581					

Variants in bold were genotyped in additional individuals
[1] Variants genotyped in 100 Xhosa individuals to confirm frequencies
[2] Variants genotyped in 50 Xhosa individuals to determine LD

4.1.3 Confirmation of the Frequencies of Prioritised Variants

RFLP analysis of all genotyped SNPs was successful (refer to Appendix 7 for RFLP gels) and all variants remained in HWE after increasing the sample size. The confirmed frequencies after RFLP analysis of an additional 85 Xhosa individuals in the previously characterised *CYP2C19*15*, *CYP2C19*9* *CYP2C19*2*, *CYP2C19*17* alleles, were revealed to be 0.09, 0.09, 0.21 and 0.10 respectively, while the novel V374I variant was detected at a frequency of 0.01. As mentioned in Section 4.1.2, variations of *CYP2C19*9* and *CYP2C19*15*, with additional variants occurring in perfect LD, were detected in this population. Confirmation of this LD was obtained through the genotyping of a total of 50 healthy Xhosa individuals. Genotyping of all 100 Xhosa individuals did not detect any of the non-functional *CYP2C19*3*, *CYP2C19*4* or *CYP2C19*7* variants in this population. Furthermore, none of the *CYP2C19*2* variants were mistakenly genotyped as a result of the adjacent *CYP2C19*10* SNP. The data for all variants which underwent RFLP analysis are shown in Table 4.1.

4.1.4 TaqMan® Copy Number Assays

A TaqMan® CNV Assay (Applied Biosystems™, Foster City, California, USA) was performed on the entire cohort of 100 healthy Xhosa individuals. To determine the number of *CYP2C19* copies present in each individual, four TaqMan® CNV Assay replicates were performed for each sample. However, to ensure that results obtained for each sample were accurate, any samples showing discrepancies in the results obtained were flagged and the four replicate assays were repeated for these samples. Samples that were flagged for repeat analysis were those samples for which insufficient amplification was obtained, those samples that exhibited a ΔC_T standard deviation > 0.21 and those samples that displayed a calculated copy number between integers (1.4 to 1.6). Additionally, analysis was repeated for any samples with a predicted *CYP2C19* copy number deviating from the expected two copies. After all samples were analysed, including those for which repeat analysis was performed, all 100 genotyped Xhosa individuals were predicted to possess two copies of *CYP2C19*, with the exception of one sample for which repeated four replicate TaqMan CNV Assays predicted only one copy of the *CYP2C19* gene. Figure 4.3A demonstrates how the reference and *CYP2C19* probes detect more or less the same level of amplification, suggesting that there are two copies of both the reference gene as well as *CYP2C19* in this sample. However, in Figure 4.3B, the amplification detected by the *CYP2C19* probe is lower than that detected by the reference probe, thereby suggesting that in this sample only one copy of *CYP2C19* is present. Figure 4.4 depicts the predicted *CYP2C19* CNV in a sample of Xhosa individuals tested by means of TaqMan® CNV Assays.

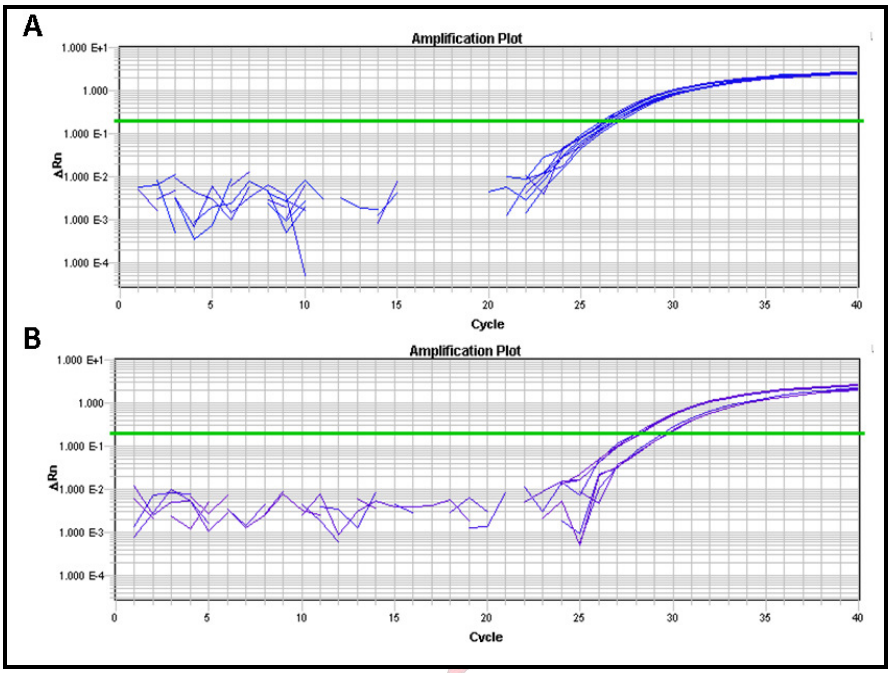


Figure 4.3: The amplification plots given for the reference and *CYP2C19* amplicons. In this Figure, A represents an individual predicted to have two copies of *CYP2C19* and B represents an individual predicted to possess only one copy of *CYP2C19.*

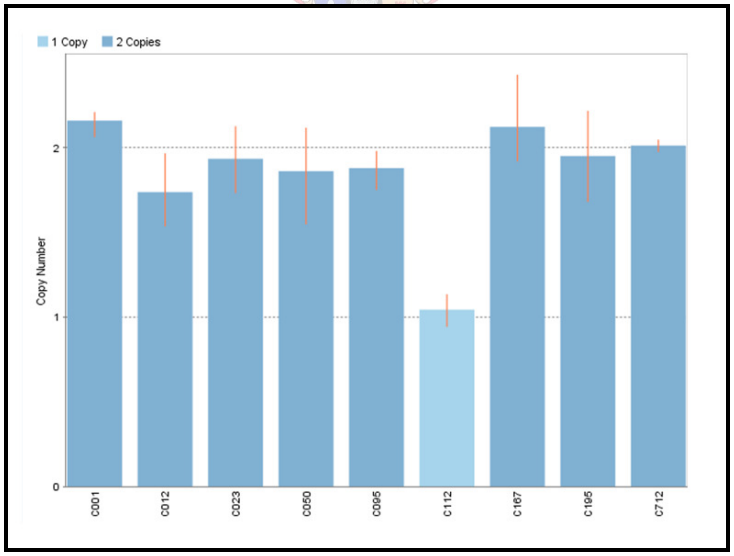


Figure 4.4: Predicted copy numbers for a sample of the Xhosa individuals examined.
In this Figure, sample C112 is predicted to possess only one copy of *CYP2C19*.

Closer inspection of the sample predicted to possess one copy of *CYP2C19* showed this sample to be heterozygous for two SNPs previously detected using RFLP analysis. Therefore the presence of a deletion in this sample was questioned. To verify the heterozygosity of the SNPs, additional bi-directional sequencing reactions were performed for this sample. The sequencing results were identical to those previously obtained using the RFLP analysis, thus verifying the heterozygosity of these variants. Since the *CYP2C19* probe sequence (supplied by Applied Biosystems™, Foster City, California, USA) was available, an additional bi-directional sequencing reaction was performed for this sample in the area to which the probe hybridises, to determine if variation may occur in the probe binding region. This was not the case. It is also plausible that sequence variation may occur in the primer binding regions, however, as the primer sequences were not supplied by Applied Biosystems™ and the primers used to amplify exon 6 in this study did not span further downstream than the probe binding area, the same procedure could not be performed to determine the presence of sequence variation in the region to which the Applied Biosystems TaqMan® CNV Assay primers hybridise.

4.1.5 Classification According to Phenotype Class

The 100 Xhosa individuals were classified according to phenotype class, such that individuals with two non-functional *CYP2C19* alleles (*CYP2C19*2*) were classified as PMs, individuals with one non-functional (*CYP2C19*2*) or two decreased function alleles (*CYP2C19*9*) were classified as IMs, individuals with two functional alleles were classified as EMs and individuals with increased function alleles (*CYP2C219*17*), unaccompanied by non-functional (*CYP2C19*2*) or decreased function genes (*CYP2C19*9*) were classified as UMs. This classification system revealed that PMs, IMs, EMs and UMs occurred at a frequency of 0.03, 0.49, 0.39 and 0.09, respectively in the Xhosa population (refer to Figure 4.5). It should be noted that none of these classes take any of the novel/uncharacterised variants detected in this population into account and that all novel/uncharacterised alleles were regarded as functional.

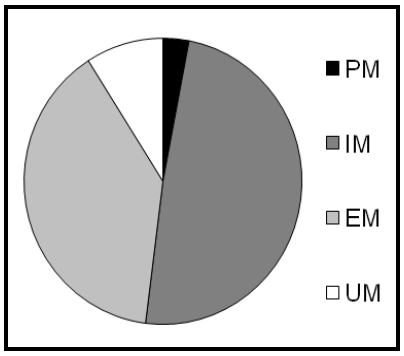


Figure 4.5: The percentage of each class of metaboliser present in the Xhosa cohort examined

4.2 Functional Analysis

4.2.1 Identification of Variants in Previously Unsequenced Area

Three additional variants were detected in the previously unsequenced 5'-upstream region, namely rs17878739, rs3814637 and rs7902257. In order to determine the alleles on which the additional variants occurred, the entire 5' region of interest was amplified, cloned into the pGL4.10 vector and sequenced. In this manner it could be determined whether the correct construct was present in the vector, as well as whether the additional variants were in fact in phase with the variants of interest. The variants identified were also genotyped in an additional 50 individuals to determine their level of LD to the previously detected variants. This RFLP analysis revealed that rs17878739 occurred in perfect LD with the *CYP2C19*15* variant (LOD = 11.7, r^2 = 1), as depicted in Figure 4.6. This figure is the same as Figure 4.2; however the additional data obtained from the variants detected through the sequencing of the unsequenced area depicted in Figure 3.2 have been added. Sequence analysis of the clones revealed that both rs17878739 and rs3814637 occurred in phase with -2030C>T and -2020C>A for the sample analysed. The frequencies and positions of the additional variants can be viewed in Table 4.1, while the chromatograms and RFLP gel images are shown in Appendix 7.

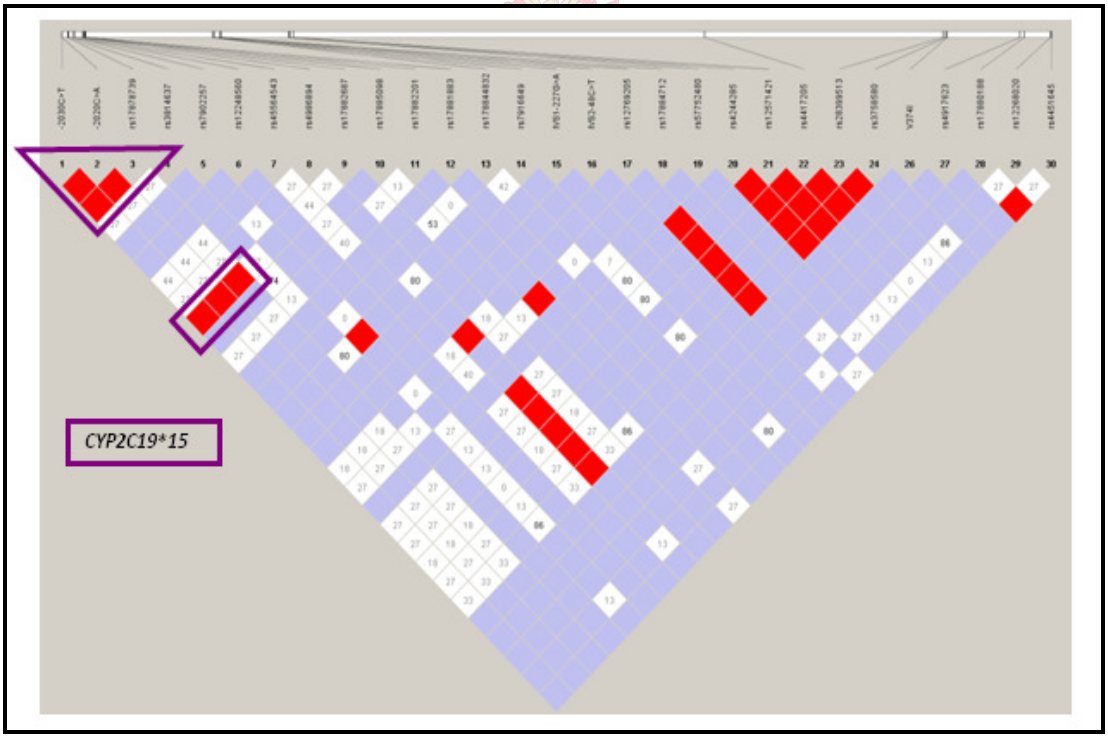


Figure 4.6: Haplotype analysis with the three additional variants detected and resultant change in the *CYP2C19*15* allele.

4.2.2 Dual Reporter Luciferase Assays

The fragments that were inserted into the pGL4.10 vectors were fragment A, containing the variants under inspection, fragment B, which is identical to the *CYP2C19* sequence as described by NCBI (NM 000769) and fragment C, which is identical to the *CYP2C19* sequence as described by Ensembl (ENSG00000165841). The constructs created through the insertion of these fragments into vectors are named according to the fragment that is inserted, such that construct A contains fragment A etc. All the fragments isolated for insertion into the vectors ranged from the nucleotide directly upstream from the ATG start codon to 2 095 bp upstream from the ATG start codon. Although successful transformation reactions were performed, sequence confirmation revealed that many of the amplified fragments that were cloned into the vectors, contained sequence errors, most likely due to the additional nested PCR reaction that was performed, therefore allowing more opportunity for the incorporation of errors. However, after sequence confirmation of a number of samples for each fragment, "error-free" constructs were identified for each desired fragment and transfection of these constructs into HepG2 cells and subsequent dual reporter luciferase assays were successfully performed. After normalisation of the obtained results, calculation of the mean, standard deviation, standard error and exclusion of outliers (all standardised values >2 or <-2) (refer to Table 4.5), differences in luciferase activity could be determined. To allow for a greater sample size, the results from all three experiments were pooled before statistical analysis was performed (refer to Figure 4.7). The Shapiro-Wilk test showed that after the outliers were removed, all samples were normally distributed (Construct A: P = 0.2406, Construct B: P = 0.1900, Construct C: P = 0.2276). The F-test demonstrated that a *t*-test for unequal variances was required when determining the difference in fold induction between construct A and C (P = 0.0123), while a *t*-test for equal variances was required when determining the difference in fold induction between construct A and B (P = 0.0939).

Although a trend towards decreased expression was observed when comparing the luciferase activity of constructs C and A, and the fold induction of construct A was a 1.5X higher than that of construct C, no statistically significant difference was observed (P = 0.0928). The difference in luciferase activity observed between constructs A and B, was however significant, with construct A exhibiting a 2.13 higher fold induction (P = 0.0077). This allele, containing the -1041G>A variant, was designated *CYP2C19*27* by the Human *CYP2C19* Allele Nomenclature website. Although, for the analysis used in this study, outliers were removed, analysis without the removal of outliers, using a Mann-Whitney U test (Mann and Witney 1947), which does not require normality, also predicted a significant decrease in the fold induction observed between constructs A and B (P = 0.004) and no significant difference between constructs A and C (P = 0.1903).

Table 4.5: Values obtained from dual reporter luciferase assays

Construct	Experiment	Normalised Value	Mean (Experiment)	Mean (Construct)	Standardised Value
Construct A	Experiment 1	0.000286814	0.000403	0.000580036	-0.518041558
		0.000509316			-0.374700514
		0.000412327			-0.437182773
	Experiment 2	0.000718096	0.000749		-0.240199645
		0.000435682			-0.422137217
		0.00109453			0.002308045
	Experiment 3	0.000761764	0.002121		-0.212067852
		0.000421756			-0.431108774
		0.005178244*			2.633130289*
Construct B	Experiment 1	0.000541339	0.000382	0.000272266	1.942324929
		0.000349176			0.555180571
		0.000254837			-0.125813812
	Experiment 2	0.000425624	0.00027		1.107025699
		0.00020635			-0.475819783
		0.000176575			-0.690754773
	Experiment 3	0.000200633	0.000165		-0.51708778
		0.000117615			-1.116360428
		0.000178246			-0.678694624
Construct C	Experiment 1	0.000417087	0.000479	0.000394429	-0.176426118
		0.00057781			0.674877965
		0.000442906			-0.039671287
	Experiment 2	0.000325764	0.000323		-0.660139906
		0.000298683			-0.803578251
		0.000345936			-0.553291729
	Experiment 3	0.000898131*	0.000548		2.371535384*
		0.000428546			-0.115730478
		0.000318696			-0.697575579

* outliers excluded from analysis

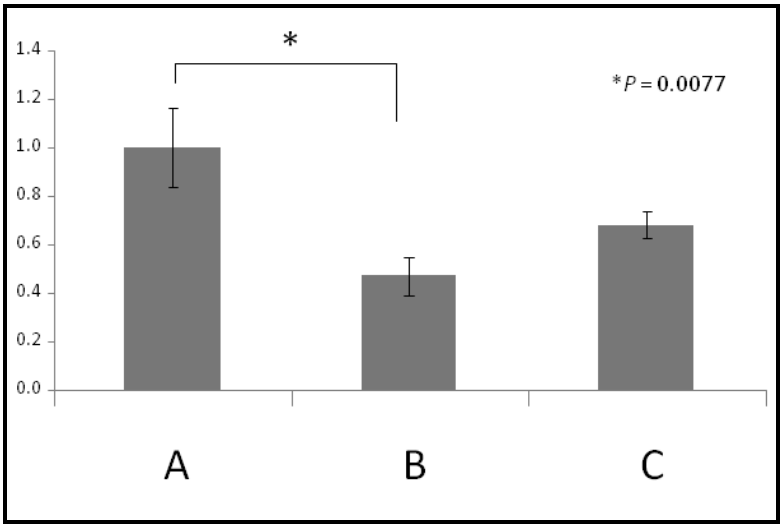


Figure 4.7: Fold induction ± SEM.
A = NM 000769 (NCBI reference sequence), B = ENSG00000165841 (Ensembl reference sequence), C = mutant (containing -2030T and -2020A variants).

The variant present in construct B, namely the rs7902257 variant, was subsequently genotyped in an additional 50 Xhosa individuals to determine the frequency of this variant in 100 individuals and was found to be 0.33. The variant was then re-analysed with the use of the PATCH, MATCH and MatInspector software programs, however, the stringency of the parameters used was decreased to match the set parameters of the various software programs. The results from this analysis are shown in Table 4.6. As it has been reported that the deletion of this region of the 5-upstream area has no impact on the expression of CYP2C19 (Arefayene *et al.* 2003), only the creation of transcription factor binding sites was taken into consideration.

Table 4.6: Predicted transcription factor binding sites created as a result of the -1041A variant

Program	**Matrix identifier**	**Core match**	**Matrix match**	**Factor name**	**Sequence**
MATCH	F$QA1F_01	0.822	0.721	qa-1F	tacactgagc**a**ttTCCCCtc
	F$ABAA_01	0.8	0.781	AbaA	cactgagC**A**TTTcccctct
	V$OCT1_02	0.814	0.743	oct-1	actgaGC**A**TTtcccc
	N$SKN1_02	0.8	0.755	Skn-1	actgAGC**A**Tttc
	V$ELK1_01	0.775	0.742	Elk-1	gagc**a**TTTCCcctctg
	V$BRN2_01	1	0.779	Brn-2	gagc**A**TTTCcctctg
	V$NFKB_C	0.973	0.713	NF-kappaB	gagc**a**tTTCCCc
	V$IK3_01	0.778	0.727	Ik-3	agc**a**TTTCCctc
	V$NKX25_01	0.783	0.705	Nkx2-5	C**A**TTTcc
Program	**Identifier**	**Mismatches**	**Score**	**Binding Factor**	**Sequence (Search Pattern)**
PATCH	RAT$VEGF_02	0	100	ER-alpha, ER-beta	GAGC**A**
	DROME$TWI_07	1	88.89	Dl	GC**A**TTTTCCC
	HS$GMCSF_03	0	100	YY1	C**A**TTT
	HS$INS_05	0	100		GGAAA**T**[1]
Program	**Matrix**	**Core similarity**	**Matrix similarity**	**Detailed Family Information**	**Sequence**
MatInspector	V$SPI1_PU1.02	1	0.96	Human and murine ETS1 factors	tgcagaggGGAAa**t**gctcagt[1]

Nucleotides underlined and in bold are those nucleotides affected by the rs7902257 variant, [1] reverse complement sequences

4.3 *In Silico* Analysis of the 5'-upstream Region

4.3.1 Comparative Sequence Analysis of the 5'-upstream Region

The high level of sequence similarity observed between all the *CYP2C Homo sapiens* and *Pan troglodytes* 5'-upstream regions did not allow for identification of individual regions of interest (refer to Figure 4.8); therefore these sequences were removed from the comparisons. Subsequent comparative sequence analysis using rVista identified eight regions of interest exhibiting greater than 75% sequence similarity to various orthologues. These sequences were used to create consensus sequences for re-alignment to the entire 5'-upstream region of each orthologue (refer to Table 3.7 for the list of orthologues used). This re-alignment of the eight consensus sequences prioritised four

regions which were shown to align to more than two species, excluding the *Pan troglodytes*. These regions will be referred to as consensus sequence 1-4 from this point on. The degree of sequence similarity observed between these regions and the human *CYP2C19* 5'-upstream sequence was subsequently determined and it was established that although the sequences aligned at different positions within the 5'-upstream regions in the different species, they all occurred within regions of LINES belonging to the L1 class (refer to Table 4.7). Subsequent transcription factor analysis showed at least two putative transcription factor binding sites in each of these regions within the *Homo sapiens* (refer to Table 4.8).

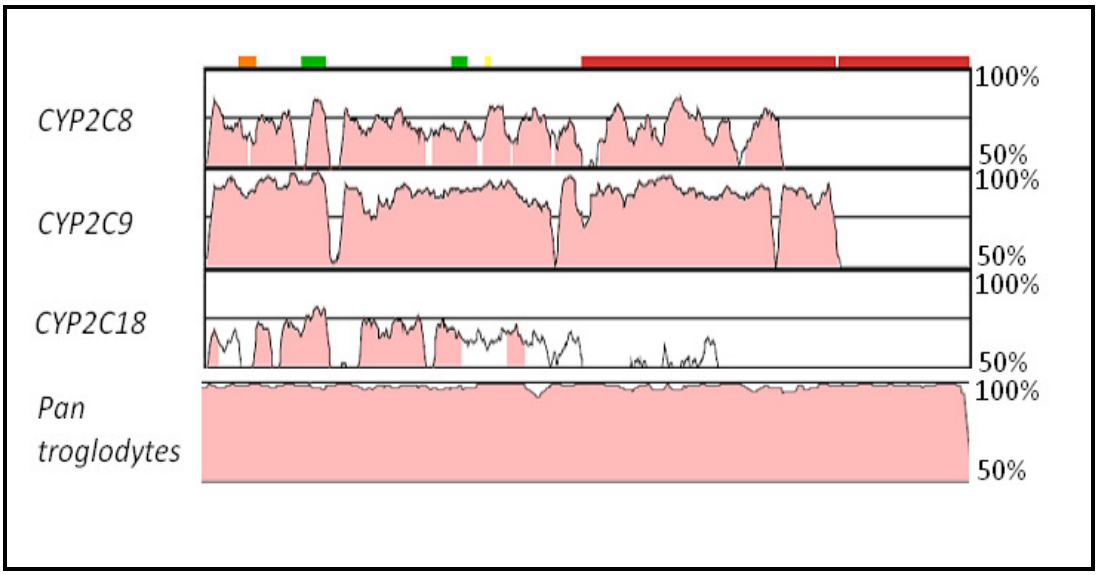


Figure 4.8: High sequence similarity observed between the *CYP2C19* reference sequence and the *Homo sapiens* and *Pan troglodytes CYP2C* 5'-upstream regions, after rVista comparative sequence analysis.

Table 4.7: Regions of high sequence similarity to the *Homo sapiens CYP2C19* 5' region

Consensus sequence 1 (133bp)

Species	Percentage identity to *Homo sapiens*	Start position (bp upstream from ATG)	Repetitive element
Homo sapiens	100	-7 164	L1PA4_3end
Macaca mulatta	87.1	-24 745	L1PA5
Pan troglodytes	86.2	-7 176	L1PA4
Mus musculus	82.4	-43 487	L1Md_A
Canis famililiaris	80.2	-3 983	L1_carn2
Rattus norvegicus	78.2	-41 877	L1_Rn2
Equus caballus	77.3	-48 217	L1-4_EC

Consensus sequence 2 (122bp)

Species	Percentage identity to *Homo sapiens*	Start position (bp upstream from ATG)	Repetitive element
Homo sapiens	100	-6 995	L1PA4_3end
Pan troglodytes	83.2	-6 998	L1PA4
Canis famililiaris	82.9	-3 844	L1_carn2
Macaca mulatta	82.1	-17 857	L1PA5
Monodelphis domestica	79.3	-3 511	L1-1_DV
Equus caballus	74.5	-47 120	L1MAB_EC
Mus musculus	70	-43 362	L1Md_A

Consensus sequence 3 (57bp)

Species	Percentage identity to *Homo sapiens*	Start position (bp upstream from ATG)	Repetitive element
Homo sapiens	100	-6 844	L1P1_orf2
Equus caballus	90.5	-47 906	L1MAB_EC
Macaca mulatta	84	-17 686	L1PA5

Consensus sequence 4 (85bp)

Species	Percentage identity to *Homo sapiens*	Start position (bp upstream from ATG)	Repetitive element
Homo sapiens	100	-6 405	L1P1_orf2
Equus caballus	91.8	-47 460	L1MAB_EC
Pan troglodytes	87.6	-12 753	L1MA3
Canis famililiaris	82.5	-3 261	L1_carn2
Mus musculus	76.1	-42 752	L1Md_A

Table 4.8: 5' Transcription factor binding sites identified in regions of high sequence similarity

Consensus sequence 1	Consensus sequence 2	Consensus sequence 3	Consensus sequence 4
C/EBP beta	CHOP-C/EBPalpha	C/EBPbeta	CHOP-C/EBPalpha
HNF-1	NF-1	HNF-3beta	GATA-3
HNF-3beta	USF	AP-1	
USF	GATA-3		
AP-1	HNF-4alpha		
NF-1			

4.3.2 CpG Island Analysis

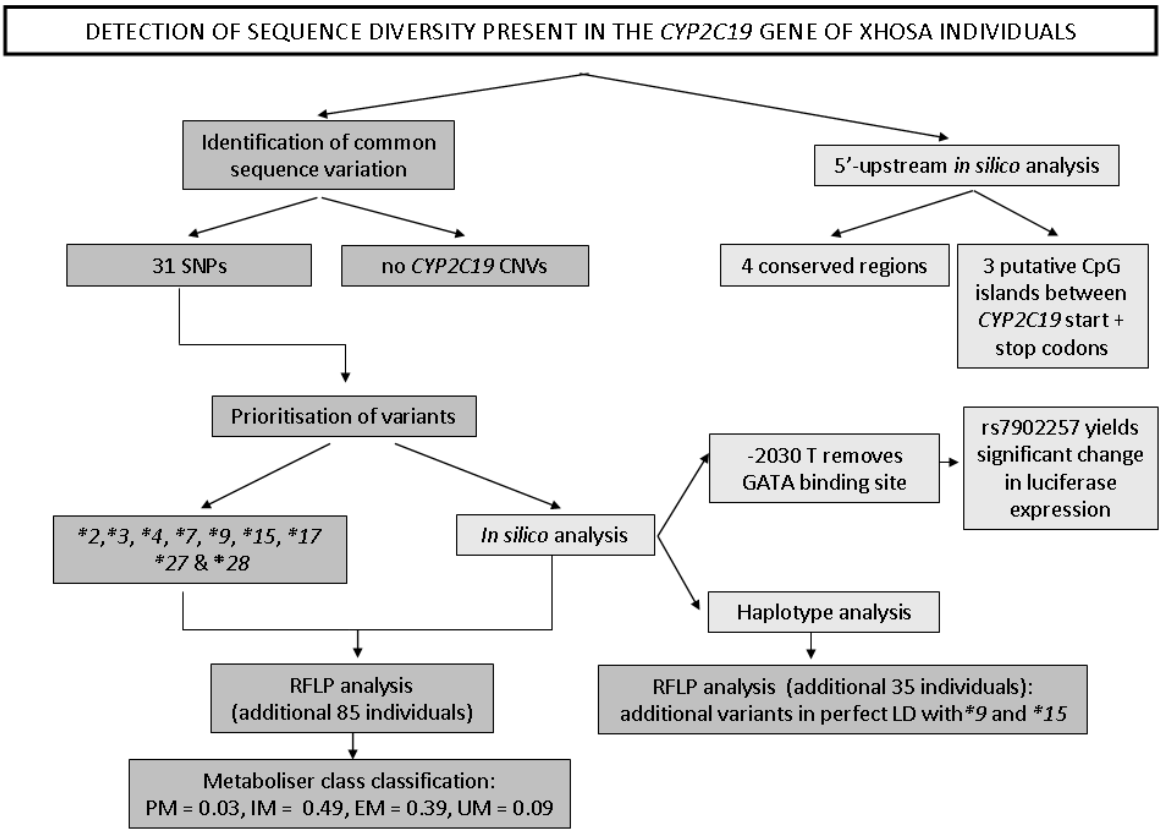
Only one CpG island occurring in the 5'-upstream region of *CYP2C8* was detected by both programs utilised. This island is, however, much closer to the downstream gene of *CYP2C8* than *CYP2C8* itself and may therefore not exert a large effect on *CYP2C8*. When considering the results obtained from the less stringent CpG plot software, more islands were detected (refer to Figure 4.9 in conjunction with Table 4.9), however, no correlation between the number of islands in and around the genes, the size of the islands and the level of *CYP2C* expression could be drawn. As the expression level of the *CYP2C* genes have been shown to vary, with the different levels of expression observed between the genes occurring in the ratio 35:60:4:1 for CYP2C8, CYP2C9, CYP2C19 and CYP2C18, respectively (Goldstein *et al.* 1994), these differences may be attributed to CpG islands. However, the only correlation observed between the CpG islands identified and the levels of expression, was that both *CYP2C18* and *CYP2C19* contain CpG islands between their translational start and stop codons, which was not observed in *CYP2C8* or *CYP2C9*. This could possibly contribute to the lower expression levels of *CYP2C18* and *CYP2C19*, when compared to the *CYP2C8* and *CYP2C9* genes.

Figure 4.9: Genomic context of the *CYP2C* genes on chromosome 10q24 (not to scale).

Table 4.9: CpG islands identified in the *CYP2C* genes

Region	Program	Number of islands	Length of islands (bp)	Position of island relative to closest ATG (bp)
CYP2C18 5'-uptream	CpG Plot	2	526 703	-20 222 -18 956
CYP2C18 (ATG-TGA)	CpG Plot	1	620	19 750
CYP2C18 3'-downstream / *CYP2C19* 5'-upstream				
CYP2C19 (ATG-TGA)	CpG Plot	3	1037 792 563	40 677 50 408 67 147
CYP2C19 3'-downstream / *CYP2C9* 5'-upstream	CpG Plot	3	575 825 671	-80 045 -58 744 -39 918
CYP2C9 (ATG-TGA)				
CYP2C9 3'-downstream / *CYP2C8 3*'-downstream	CpG Plot	2	578 757	33 925 44 180
CYP2C8 (ATG-TGA)				
CYP2C8 5'-upstream	CpG Plot	5	579 565 652 867 933	-41 423 -52 471 -66 801 -100 915 -113 923
	CpG island searcher	1	1214	-113 715

4.4 Summary of Results



CHAPTER 5:

DISCUSSION

CHAPTER 5: DISCUSSION

5.1 The Xhosa Population Under Comparison

This study detected a high level of genetic variation in the Xhosa population, which is consistent with the level of genetic variation expected in African populations (Tishkoff *et al.* 2009). This finding also correlates with the data obtained from a study performed by Blaisdell *et al.* (2002), where a high level of *CYP2C19* variation was detected in African populations when compared to the other populations examined. The re-sequencing of *CYP2C19* in 24 individuals from the African (African American and African Pygmies), Asian and Caucasian populations revealed 26, 18 and 14 variants, respectively (Blaisdell *et al.* 2002). Recently, it has been suggested that when focusing more specifically on African populations, Southern African populations are expected to show greater genetic diversity (Tishkoff *et al.* 2009) (refer to Figure 5.1). This may be indicative of the ancient origins of populations residing in Southern Africa. The high level of genetic diversity present in Southern African populations is further validated by comparing the sequencing data obtained from the Central African and African American individuals in the Blaisdell *et al.* (2002) study, to the data obtained in this study, where more variants (30 variants vs 26 variants) were discovered in the Xhosa population through the re-sequencing of fewer individuals (15 individuals vs 24 individuals). It must, however, be considered that the small sizes in both studies may contribute to a greater likelihood of the differences observed to occur as a result of chance.

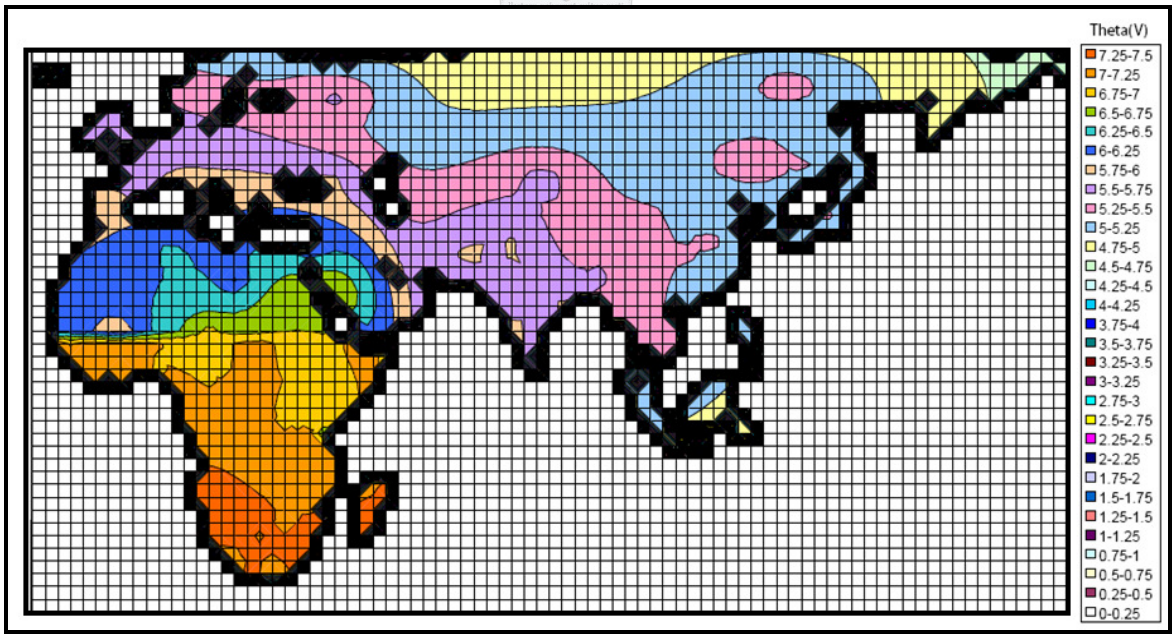


Figure 5.1: Genetic diversity based on variance in microsatellite length.
The greatest variance in microsatellite length is depicted in orange in the region of the Southern African populations. Further north of this point the variance gradually decreases.
(Tishkoff *et al.* 2009) (Reprinted with permission from The American Association for the Advancement of Science)

The differences observed between African populations, including the South African Xhosa population and other more frequently studied African populations such as the HapMap 3 populations, namely the Nigerian, Kenyan and African American populations emphasize the need for genomic information, such as the data obtained by this study, to be generated for Southern African populations. Figure 5.2 shows how the admixture (depicted by the different colours) from populations residing in Southern Africa, differs substantially from the admixture observed in other populations residing within the same continent, therefore emphasising that each African population may represent an independent entity, with for example the South African Xhosa population exhibiting a genetic profile distinct from other more northern African populations.

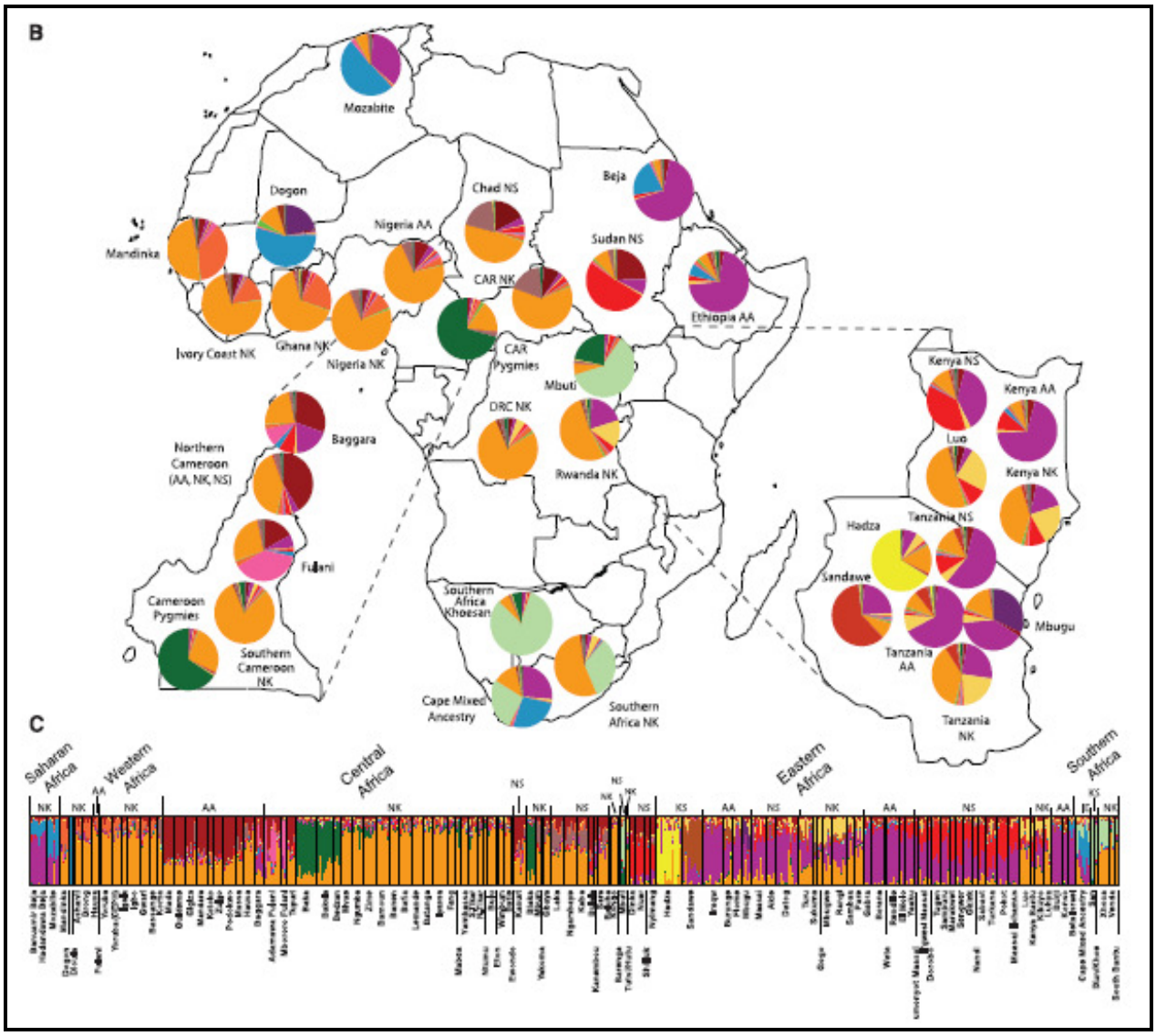


Figure 5.2: The admixture observed in the different African populations, where Southern African populations appear to differ quite substantially from other African populations.
(Tishkoff *et al.* 2009) (Reprinted with permission from The American Association for the Advancement of Science)

Through various projects, such as the 1000 Genomes Project, vast amounts of genomic data are being generated to aid future studies. Unfortunately, this data remains limited for Southern African populations. The 1000 Genomes Project aims to obtain the genomic information from approximately 1 200 individuals from different populations around the world (http://www.1000genomes.org/page.php?page=home). In the words of Dr Durbin, the co-chair of the 1000 Genomes Project, "The 1000 Genomes Project will examine the human genome at a level of detail that no one has done before." However, although this project will examine a wide range of populations (refer to Figure 5.3), including those of African origin, it shall not examine any Southern African populations. This lack of data does put these populations at a disadvantage in some instances; nonetheless it provides many opportunities for research in South Africa. It is important that we appreciate the wealth of information that remains to be elucidated from South African populations. For this reason it is crucial that the genetic information available from these populations is comprehensively utilised and protected by local institutions. To protect the information obtained from the DNA of South African populations, a set of appropriate ethics guidelines need to be established and followed, so as to ensure that these local populations are not exploited.

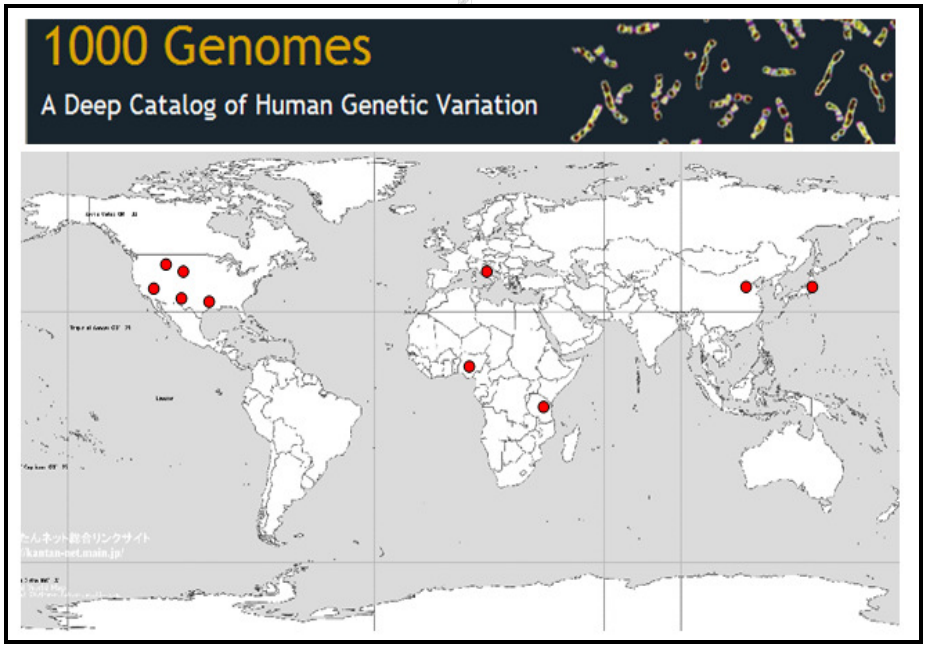


Figure 5.3: The populations to be sequenced by the 1000 genomes project.
Red dots depict the areas from which the populations to be sequenced have been taken. Although African populations are represented, there are no samples representing Southern Africa (http://www.1000genomes.org/page.php)

When focussing specifically on the Xhosa population, the comparisons made between the presence and minor allele frequency (MAF) of the variants detected in the Xhosa (refer to Table 4.1) and other populations (Blaisdell *et al.* 2002), revealed some interesting observations (refer to Figure 5.4). When compared to the Asian and Caucasian populations, the African populations (Blaisdell *et al.* 2002), showed a larger number of variants, some of which occurred at lower frequencies, indicative of the older, more diverse structure of these populations. In contrast, the Asian and Caucasian populations showed fewer variants, with a higher frequency of certain variants, indicative of the bottleneck effects that occurred in these populations during their migration out of Africa. Interestingly, the Xhosa population showed the greatest number of variants; however, these variants were present at a relatively high frequency. The high level of variation present in the Xhosa population may be attributed in part to admixture with the Khoisan population (Tishkoff *et al.* 2009). The Khoisan have been shown to form the root of the *Homo sapiens* mitochondrial tree, suggesting that they are the most ancient of populations (Gonder *et al.* 2007), hence the high level of genetic variation in this population. On the other hand, the high frequency of variants observed may be attributed to the bottleneck effect that the Xhosa population experienced as a result of the Xhosa cattle-killing of 1856-1857. This crisis occurred as a result of a message relayed by the Xhosa prophetess, Nongqawuse, who prophesised that if the Xhosa destroyed all their cattle and crops, the spirits of the ancestors would destroy the white settlers. The loss of livelihood and food resulted in the Xhosa population residing in British Eastern Cape, decreasing in number from approximately 105 000 to 25 916 individuals (Peires 1989), representing the loss of approximately three-quarters of this population. Thus, the Xhosa population present an interesting population to study as they may exhibit remnants of both an ancient genome as well as a recent bottlenecking.

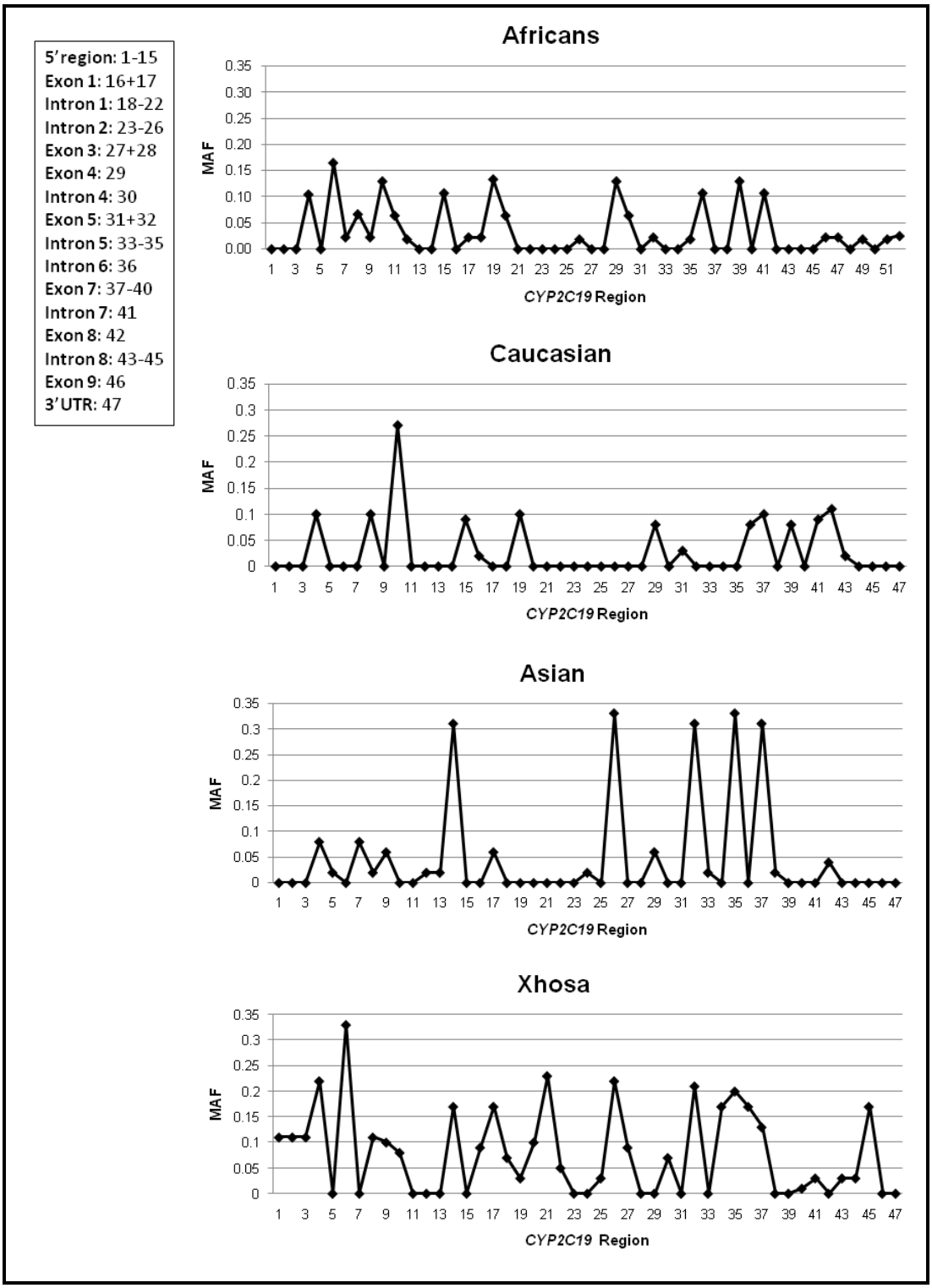


Figure 5.4: The MAF vs. *CYP2C19* region of four different populations.
Each dot represents a SNP that was detected in at least one of the populations, with those dots falling on the baseline representing the SNPs that were not detected in the specific population.

5.2 Variants Observed in this Study

In section 5.1 it was noted that the re-sequencing of the *CYP2C19* gene in 15 South African Xhosa individuals revealed a high level of sequence variation. This re-sequencing analysis revealed a total of 30 variants (refer to Table 4.1), including those that were detected after the initially unsequenced 5'-upstream area was analysed. For discussion purposes these variants have been ordered into the following categories: i) previously identified *CYP2C19* alleles (http://www.cypalleles.ki.se/cyp2c19.htm), ii) variants not previously reported on the Human *CYP2C19* Allele Nomenclature website, but which have been previously reported (http://www.ncbi.nlm.nih.gov/sites/entrez), iii) novel variants and iv) copy number variants.

5.2.1 Previously Described Human *CYP2C19* Alleles

The variants detected in the Xhosa population which have been previously identified and named on the Human *CYP* Allele Nomenclature website (http://www.cypalleles.ki.se/cyp2c19.htm) are the allele-defining variants for *CYP2C19*2*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17*. Each of these variants represent CYP2C19 proteins with different functionalities, such that *CYP2C19*2* results in a non-functional enzyme, *CYP2C19*9* results in a decreased function enzyme, *CYP2C19*15* seemingly has no effect on enzyme functioning and *CYP2C19*17* results in an increased expression of the enzyme. This range of enzyme functionalities in the population emphasises the diversity present in African populations. It is important to note here, that although *CYP2C19*2* and *CYP2C19*17* have been studied *in vivo* (Furuta *et al.* 1999; Sagar *et al.* 2000; Shirai *et al.* 2001; Sapone *et al.* 2003; Gardiner and Begg 2006; Schroth *et al.* 2007; Shuldiner *et al.* 2009), the data for the other two variants are inconclusive. Although convincing *in vitro* evidence is available for *CYP2C19*9* (Blaisdell *et al.* 2002), no known phenotypic validation has been performed. Furthermore, *CYP2C19*15*, by virtue of its proximity to the N-terminus, has been assumed to have no effect on the enzyme functioning (Blaisdell *et al.* 2002), therefore no functional validation of this variant is available. This highlights a need for genotype-phenotype correlations to be performed for these alleles. Additionally, as it cannot be assumed that the phenotypic data obtained for the *CYP2C19*2* and *CYP2C19*17* alleles in Caucasian populations will be identical to that observed in the Xhosa population; genotype-phenotype correlations for these variants are still required in the Xhosa population. It is notable that the variants making up these alleles, as reported on the Human *CYP* Allele Nomenclature website, are not exactly the same as those observed in the Xhosa population examined in this study. This is however to be expected considering that the LD observed is likely to differ from population to population (Tenesa *et al.* 2007).

5.2.2 Functional Validation of an Uncharacterised Variant

The uncharacterised variants identified in this population were those variants which have been assigned rs numbers, but have not been previously described on the Human *CYP* Allele Nomenclature website. None of these variants were predicted through the various *in silico* analyses utilised, to exert an effect on CYP2C19. However, as one of these uncharacterised variants, namely the rs7902257 variant, was present in the Ensembl sequence (ENSG00000165841) used as a reference sequence throughout, it was decided to include this variant in a construct for the dual reporter luciferase assay experiments (refer to Figure 3.3).

Unexpectedly, the results from the dual luciferase reporter assays revealed that the presence of this variant did result in a significant decrease in the fold induction observed ($P = 0.0077$) (refer to Figure 4.7) when compared to the construct containing the NCBI reference sequence (NM 000769). After re-evaluation by the transcription factor binding site prediction algorithms, using less stringent parameters, several transcription factor binding sites were predicted to be created or removed as a result of this variant (refer to Table 4.6). As this area has been deleted without effect on CYP2C19 expression (Arefayene *et al.* 2003), only sites that were created as a result of this variant were evaluated. Additionally these predicted sites were required to suppress activity in order to verify the decrease in expression observed in the luciferase assays. Although none of the programs used predicted the creation of the same sites, two promising candidates were identified by the MATCH prediction algorithm. These two candidates were the octamer binding protein-1 (Oct-1) and the nuclear factor kappa-B (NF-kappaB), both of which have been implicated in the decreased expression of *CYP* genes (Thum *et al.* 2000; Lee *et al.* 2000; Assenat *et al.* 2006; Fiala-Beer *et al.* 2007). Oct-1 has been reported to repress *CYP1A1* by binding to a site 0.8 kb upstream from the transcriptional start site (Bhat *et al.* 1996). Additionally, there is evidence for an Oct-1 binding site in *CYP4A2* (Fiala-Beer *et al.* 2007). Although there is strong evidence for NF-κB repressing the expression of *CYP2C* genes, the mechanism by which this is achieved, seems to point towards the inhibition of PXR and CAR expression, thereby indirectly decreasing *CYP2C* expression, as opposed to binding directly to the promoter sequence (Assenat *et al.* 2006). Since Oct-1 has been reported to act directly on the *CYP2C* promoter region, the creation of this transcription factor binding site as a result of the rs7902257 variant in the *CYP2C19* 5' region, seems more likely to influence the decrease in transcription of *CYP2C19* in this case.

The results obtained from the rs7902257 variant (*CYP2C19*27*) serve as a reminder that the parameters used in *in silico* analysis should not be overly stringent, as important information may be missed as a result. Although the dual luciferase reporter assays used predicted a significant decrease

in luciferase activity as a result of this variant, once again *in vivo* verification of the results obtained are required to verify the functionality of this variant.

5.2.3 Novel Variants and Functional Verification

To our knowledge, this study detected five variants which have not been reported in any other studies, which are referred to as novel variants throughout this thesis (refer to Table 4.1). The various *in silico* analyses performed on these five novel variants predicted three of the five novel variants to exert an effect on CYP2C19. In order to consider the strength of these predictions, it may be important to consider the putative nature of bioinformatic techniques. Bioinformatics has immense power to prioritise vast amounts of data as well as to predict the effect of this data. However, unfortunately as of yet, the complexity of biological systems is not yet fully understood, therefore these programs are unable to compensate completely for experimental analyses. The bioinformatic analysis used in this study did, however, serve to correctly predict the effect of previously characterised variants such as the allele-defining *CYP2C19*2* and *CYP2C19*7* variants (refer to Table 4.2); as such the results obtained from the *in silico* analysis for the novel variants appear to be reliable. Even so, the effects of these variants on CYP2C19 do require further experimental validation.

The first of the novel variants that will be discussed is the V374I variant (*CYP2C19*28*) found in exon 7. This variant results in a change in the amino acid code, namely a change from valine to isoleucine, and thus the protein composition. However, due to the similar nature of valine and isoleucine, further *in silico* analysis did not reveal a change in protein structure. Furthermore, after reviewing the literature, no data could be found suggesting that this amino acid is involved in the active site of CYP2C19. It is, however, essential that no variation is discarded without thorough analysis, as even synonymous variants may affect the level of protein expressed. It has been reported that if a variant results in a change in the secondary structure of the mRNA, this may influence the binding of ribosomal subunits and therefore increase the exposure of the mRNA to nucleases, which will degrade the mRNA (Kudla *et al.* 2009). Although the mFold analysis performed only showed a slight change in the secondary structure (refer to Figure 5.5), the seeming insignificance of a variant should never be assumed. As it is impossible to know with absolute certainty what the effect is that this variant exerts on the protein, phenotyping studies are recommended.

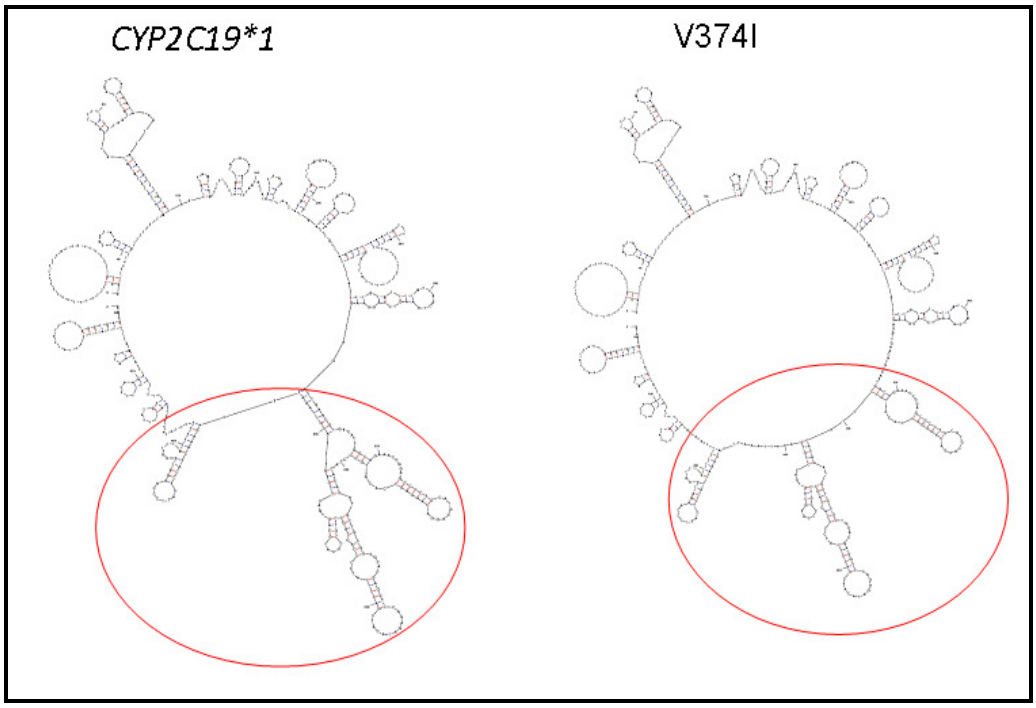


Figure 5.5: Predicted differences in mRNA folding observed between the *CYP2C19*1* allele and V374I, predicted by mFold analysis.

The other two novel variants predicted to exert an effect on CYP2C19, were the –2030T and -2020A variants, which were expected to influence the transcription of the gene. All three the transcription factor binding site prediction algorithms convincingly suggested the removal of a GATA factor binding site (refer to Table 4.3). According to the literature, GATA plays a role in the expression of *CYP* genes (Thum *et al.* 2000). Additionally Sim *et al.* (2006) initially suspected GATA factors to exert an effect on *CYP2C19* expression, when examining the effect of *CYP2C19*17*.

For this reason, dual luciferase reporter assays were performed in order to functionally validate the predicted effect of these two novel variants. Although the results from these assays showed a trend towards a decrease in the fold induction observed as a result of these variants when compared to the construct containing the NCBI reference sequence (NM 000769), which approached significance (P = 0.0928) (refer to Figure 4.7), the data did not provide conclusive proof for decreased expression. It is important to remember that there may be several reasons for the lack of significance obtained from these results. Firstly, the predicted removal of the GATA-factor binding site is putative, hence the need for functional validation. Secondly, although it has been suggested that GATA factors may play a role in *CYP2C19* expression, there is no conclusive evidence for this. Next, the presence of the other variants occurring alongside the variants of interest (refer to Figure 3.3) could influence the expression of the gene, thereby confounding the results obtained for the dual reporter luciferase assays used in this study. To determine if this is indeed the case, further studies making use of more specific constructs, containing only the variants of interest, perhaps through the use of site-directed

mutagenesis, should be performed. Lastly, the inserted 5'-upstream region could contain regions, such as other GATA factor binding sites, which may compensate for the removal of the GATA-factor binding site as a result of these variants. Once again, as the difference in fold induction was shown to approach significance, all these factors should be considered and the effects should perhaps be re-evaluated *in vivo*, with the use of phenotype studies.

5.2.4 Copy Number Variation (CNV)

The last category of genetic variation that was examined by this study was the presence of CNV in the *CYP2C19* gene. It has been estimated that approximately 12% of the genome consists of CNVs, of which duplications as opposed to deletions are more common (Redon *et al.* 2006; Stranger *et al.* 2007). Additionally deletions have been reported in *CYP2D6*, while duplications have been reported in both *CYP2A6* and *CYP2D6* (http://www.cypalleles.ki.se/). The occurrence of CNVs in these *CYP* genes, along with the fact that *CYP* duplications have been reported to occur at a higher frequency in African populations, perhaps due the more diverse diets of African individuals (Ingelman-Sundberg *et al.* 2007), make the occurrence of CNVs in the *CYP2C19* gene plausible. Additionally, as CNVs may occur as a result of unequal crossing over, the high level of sequence similarity observed between *CYP2C19* and *CYP2C9*, which lie adjacent to each other on chromosome 10, suggest the possibility of unequal crossing over between these two genes and thus the development of *CYP2C19* deletions and duplications (refer to Figure 5.6). Although to our knowledge, no *CYP2C19* CNVs have been reported, the CNVs observed between different populations have been shown to vary substantially, with one study comparing African Americans and Caucasians, reporting that while only 72 CNVs were common to both populations, 412 and 580 were unique to the African American and Caucasian populations, respectively (McElroy *et al.* 2009). Thus, considering the under-representation of extensive *CYP2C19*-based studies in African populations, and more specifically Southern African populations, the possibility of *CYP2C19* CNVs occurring in the Xhosa population, required further investigation.

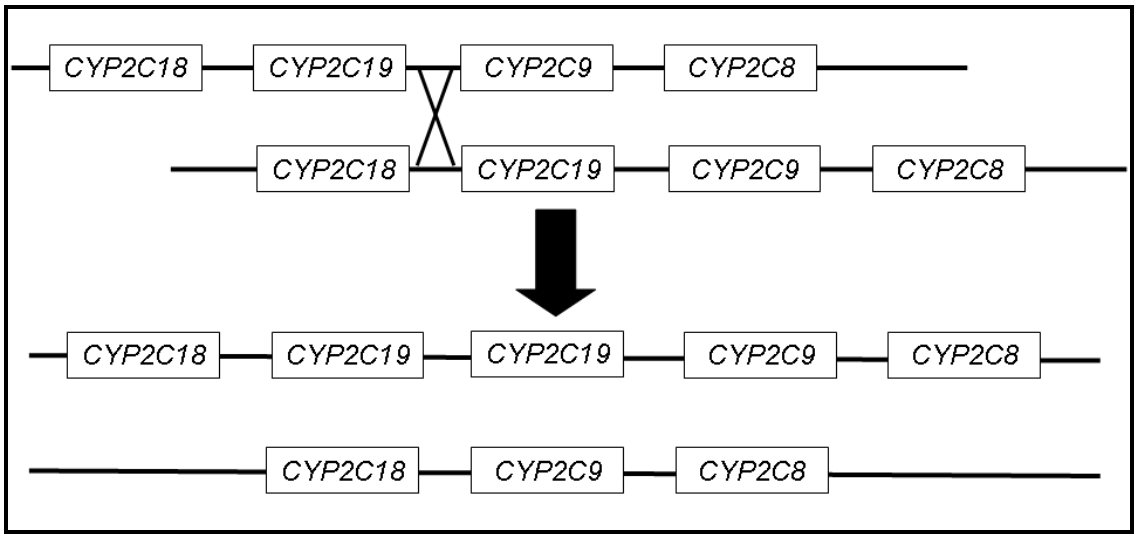


Figure 5.6: The mechanism by which *CYP2C19* duplication and deletions may occur, by which the high sequence similarity observed between *CYP2C19* and *CYP2C9* may allow for an unequal crossing over event.

CNVs ranging up to 950 824 bp have been reported on chromosome 10 (http://projects.tcag.ca/variation), however, no *CYP2C19* deletions or duplications were detected in the Xhosa cohort used in this study. This lack of *CYP2C19* CNV is not necessarily indicative of a low frequency of *CYP* CNV present in this population, as both duplications and deletions have been detected for *CYP2D6* in this Xhosa cohort (Wright *et al.* 2009, manuscript submitted to *Prog Neuropsychopharmacol Biol Psychiatry*). Although no CNV was detected in this population for *CYP2C19*, only one probe hybridising to intron 6 was utilised. Therefore the possibility of additional CNVs in other areas, such as the 5'-upstream area, cannot be ruled out. It remains possible that smaller deletions or duplications may be located within the 5'-upstream region of *CYP2C19*, where nine LINE L1 regions have been identified by Repeat Masker. These micro CNVs may occur due to a replication fork collapse, which may result under stressful conditions. This disrupted region may then align to a region of microhomology on either side of the break point, subsequently resulting in a CNV (Hastings *et al.* 2009). Regions containing LINEs and SINEs have been implicated in these micro CNVs (de Smith *et al.* 2008), possibly due to the fact that these regions are more susceptible to DNA breakage as a result of their transposon properties (Hastings *et al.* 2009). Future studies could thus examine possible 5'-upstream CNVs, which would not have been detected by the analysis used in this study.

Furthermore, it is important to bear in mind, that although the TaqMan® CNV Assay is easy to use, there are a few shortcomings. This is illustrated by the fact that one of the samples tested by this assay was predicted to possess only one copy of *CYP2C19*, despite being heterozygous for two SNPs in the gene, even when both SNP and CNV genotyping protocols were independently replicated. The heterozygosity of this sample would indicate that the TaqMan® CNV Assay prediction is incorrect and

that there are in fact two copies of the *CYP2C19* gene. The discrepancy in the results obtained, could be explained by the presence of variants within the regions where either the primers or probe were designed to bind. The variants present in these areas would influence the stable hybridisation of these oligonucleotides, resulting in a decrease in either the amplification detection or of the amplification process itself, and therefore result in an effect similar to that observed if one copy of the gene is deleted. As variation is common in *CYP* genes, especially in African populations, this is a plausible explanation. The region to which the probe binds was included in the TaqMan® CNV assay information file, therefore the presence of variation in this region could be elucidated through bi-directional sequence analysis. This analysis did not detect any SNPs within the probe binding region, however the rs4417205 variant was heterozygous in this fragment. This analysis although not providing an explanation for the TaqMan® CNV Assay, further validates the presence of two copies of the *CYP2C19* gene within this sample, due to the heterozygosity of the rs4417205 variant. As the sequences of the primers designed specifically by Applied Biosystems™ for this particular TaqMan® CNV Assay were not included in the information file, it remains plausible that variation within the regions to which these primers bind may result in an effect similar to that observed in the genotyping methods utilised by amplification refractory mutation systems (ARMS) and allele-specific-PCR (AS-PCR) analysis, where primers do not bind, as a result of a mismatch close to the 3' end (Singh *et al.* 2009). If this is indeed the case, the variation responsible for the mismatch is predicted to be rare in the Xhosa population, as only one sample out of the 100 genotyped, appears to have been affected in this manner. The combined results obtained from this analysis where the presence of the predicted *CYP2C19* deletion was negated by the heterozygosity of the sample, serve as a reminder that the data obtained from one analysis should validate and complement the data from other analyses.

5.3 *CYP2C19* Population Comparisons

5.3.1 Comparisons of the *CYP2C19* Variants Detected

This study focused on the examination of *CYP2C19* in the Xhosa population. It is, however, important that the data obtained in this study is compared to the data obtained from previous studies in order to determine to what extent the previously obtained data may be applicable to this and other South Africa populations. By doing this, it may be possible to build on the information already available and to use this information as a guideline for future studies.

It was mentioned in the literature review (refer to Section 2.3.3) that there is a large amount of information available regarding the functionality and presence of the *CYP2C19*2* and *CYP2C19*3* alleles in various populations. However, data regarding other variants appears to be less common. In order to compare the presence and frequency of the most important variants occurring in the

Xhosa population, data from four additional populations were obtained. These populations are the Caucasian, the Asian, the Cape Mixed Ancestry (CMA) and the Venda populations. As the Venda and Xhosa population both belong to Eastern-Bantu speaking populations residing in South Africa (Lane *et al.* 2002), the comparison between these populations will be considered separately, as they are expected to show the greatest level of similarity.

With regards to the CMA population, it is important to note that the analysis performed on this population was designed according to the variants detected in the Xhosa population. The variants that were genotyped by our research group in 75 individuals from the CMA population were the allele-defining variants for *CYP2C19*2*, *CYP2C19*3*, *CYP2C19*9*, *CYP2C19*15*, *CYP2C19*17* as well as V374I, -2030T and -2020A. Due to the fact that at the time of this project no results for the dual luciferase reporter assays were available, the rs7902257 variant was not genotyped in this population. It should be noted that the genotyping of this variant in the CMA population in the future is, however, recommended. The CMA present an interesting population on which to focus as it has been reported that this population exhibits the highest level of admixture worldwide (Tishkoff *et al.* 2009). It has been reported that this population arose approximately 350 years ago with the arrival of the European settlers and their slaves in the Cape. The arrival of these populations directly led to the creation of the mixed ancestry population, which has been shown to exhibit admixture from European, South Asian, Indonesian and Xhosa populations (Patterson *et al.* 2009). Therefore by comparing the data available for the Asian and Caucasian populations, in combination with the data obtained by this study for the Xhosa population, it was possible to deduce whether this reported admixture in the CMA population is observed in the *CYP2C19* gene (refer to Figure 5.7).

When examining the frequency comparisons made between these four populations, it is interesting to take note of the rs7902257, *CYP2C19*3*, *CYP2C19*17* and V374I variants. With regards to the rs7902257 variant, although not genotyped in the CMA population, this variant was shown to result in a significant decrease in the fold induction observed in the dual luciferase reporter assays performed by this study. As this variant occurs at a high frequency in the Xhosa population when compared to the Asian and Caucasian populations, the effect of this variant may be especially important in the context of pharmacogenetics in South Africa. This may serve to emphasize how certain variants affect specific populations to a greater extent than others, further pointing out that the genotypes observed in one population cannot necessarily be directly applied to another population.

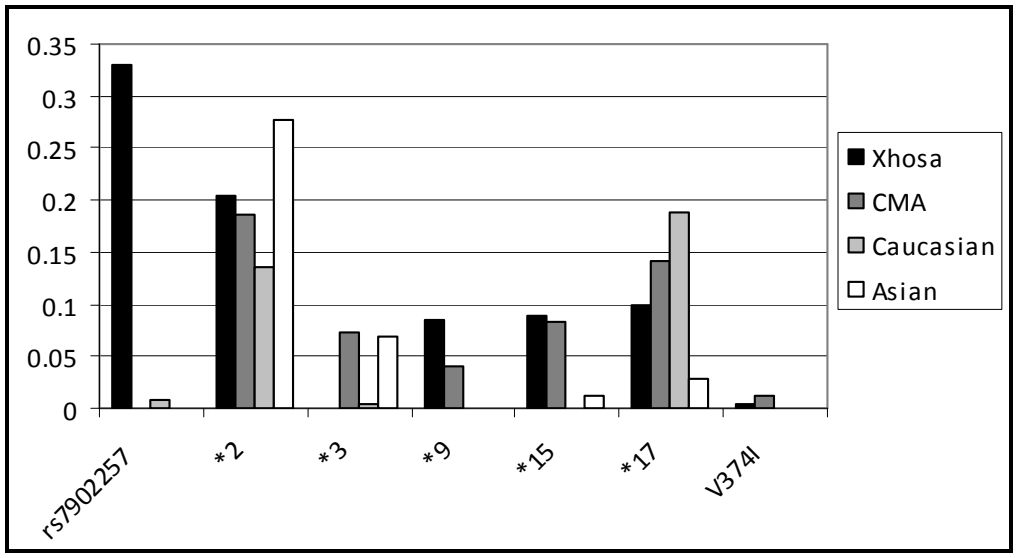


Figure 5.7: Frequency comparisons between the CMA population and various other populations. (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=7902257, Da Silva *et al.* 2009, Blaisdell *et al.* 2002, Sim *et al.* 2006, Chen *et al.* 2008, Sugimoto *et al.* 2008, Justenhoven *et al.* 2008, Ragia *et al.* 2009, Sistonen *et al.* 2009)

Focusing on the *CYP2C19*3*, *CYP2C19*17* and V347I variants, this is the first study, to our knowledge, to report the occurrence of these variants in a South African population. As the *CYP2C19*17* variant results in a class of UM for CYP2C19, this variant will be discussed at a later stage. Focussing on the *CYP2C19*3* variant, although this variant was not detected in the Xhosa population, the presence of this variant in the CMA population may be indicative of the Asian admixture present in this population. With regards to the V374I variant, it is the first time to our knowledge that this variant has been reported in any population group, therefore the presence of this variant in the CMA group may be indicative of the Xhosa admixture in this population. All other variants genotyped in the CMA population appeared to exhibit intermediate frequencies between the populations, indicating a high level of admixture observed in the CMA group, similar to that reported by Tishkoff *et al.* (2009).

Additionally, the absence of the allele-defining variants for *CYP2C19*3*, *CYP2C19*4*, *CYP2C19*7* and *CYP2C19*10* in the Xhosa population does not necessarily rule out the presence of these variants in the CMA population, as has been illustrated by the presence of the *CYP2C19*3* variant in the CMA population. The genotyping of the other non-functional variants is therefore recommended in the CMA population in the future. Determining the likely origin of each of these variants may provide clues as to the likelihood of their detection in the population. Therefore, as *CYP2C19*4* has been reported to occur in both Caucasian (Ferguson *et al.* 1998) and Asian populations (Garcia-Barceló *et al.* 1999), while *CYP2C19*7* has also been reported in Caucasian populations (Ibeanu *et al.* 1999), it is probable that these variants will occur in the CMA population. Furthermore, although *CYP2C19*10* has been reported to occur in African American individuals (Blaisdell *et al.* 2002; Rasmussen 2008),

this allele has not been detected in the Xhosa population. This emphasizes once again a need to consider the African American populations and Southern African populations as separate groups, due to the fact that a variant occurring in one population may not necessarily be indicative of its presence in another population.

When comparing the Xhosa and Venda populations, which are both Bantu-speaking populations belonging to the Niger-Kordofanian African macrofamily (Ehret 2001), exhibiting high levels of components from the Southern African Khoisan and Western African Bantu associated ancestral clusters (AAC), as well as low levels of East African Bantu AACs (Tishkoff *et al.* 2009), it is to be expected that their genetic composition should be similar, as mentioned earlier. The comparisons made between the *CYP2C19* data obtained for these two populations, although showing similarities, do display some discrepancies which may be attributed to the small sample size utilised for the examination of the Venda population (refer to Figure 5.8). Firstly, when considering the two most extensively characterised *CYP2C19* variants, namely *CYP2C19*2* and *CYP2C19*3*, the Venda and Xhosa appear to show similar tendencies. In both populations *CYP2C19*3* is absent, while *CYP2C19*2* occurs at similar frequencies (21.7% in the 75 Venda individuals (Dandara *et al.* 2001) and 20.5% in the 100 Xhosa individuals). All other comparisons made, although suggesting differences between the populations, are complicated by the small number of individuals genotyped in the Venda population (nine Venda individuals vs 100 Xhosa individuals). The inaccuracy of small sample sizes in determining frequencies, is highlighted by the discrepancies observed in the frequencies as determined in the 15 Xhosa individuals utilised for sequence analysis as opposed to the 100 Xhosa individuals genotyped with RFLP analysis in this study (refer to Table 4.1, e.g. SNPs V374I and rs17882201). Additionally, it is important to note that the analysis performed on the Venda individuals did examine beyond the -98T>C variant. It is therefore recommended that before any direct conclusions are drawn between the two populations, a more thorough analysis should be performed. It does, however, appear that if one population is thoroughly analysed, such as the Xhosa population examined in this study, the data generated may be used as a guideline for the genotyping of other Eastern-Bantu speaking populations of South Africa. Therefore the data obtained by this study on the Xhosa population may be especially useful for elucidating the *CYP2C19* genetic profile of the South African Zulu population which appear to be most closely related to the Xhosa (Lane *et al.* 2002) and are the largest population residing within South Africa, making up 23.8% of the country's population.

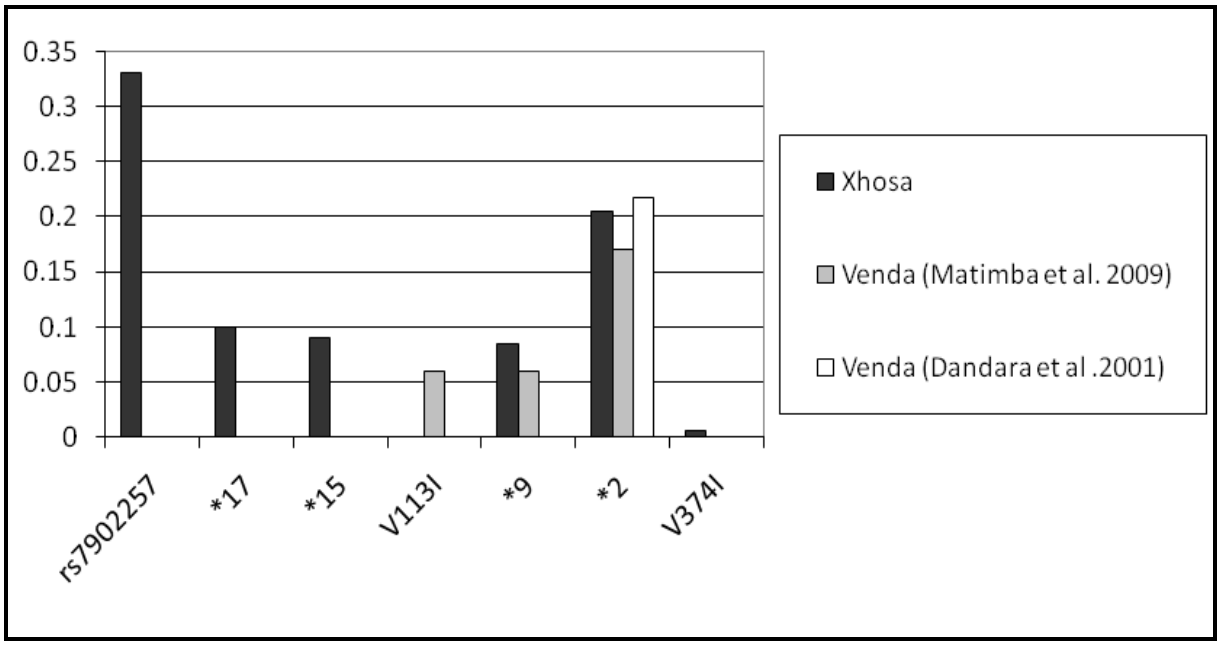


Figure 5.8: Frequency comparisons between the Venda and Xhosa populations.
For these comparisons 100 individuals were genotyped in the Xhosa population, 75 individuals were genotyped in the Venda population examined by Dandara *et al.* (2001) (examining only *2 and *3) and 9 Venda individuals were genotyped in the Matimba *et al.* (2009) study. Neither the -1041A nor *17 variants were examined in the Venda populations.

The comparisons discussed in this section, especially those comparing the South African populations, have shown that before pharmacogenetic information may be applicable to South Africa, the pharmacogenetic profiles of the different South African individuals residing in the "rainbow nation" are required, as was highlighted by the comparisons made between the Xhosa and CMA populations. Although the data obtained for the Venda and Xhosa populations showed similarities, discrepancies between the two populations were still present. Therefore the wide variety of different population groups present in South Africa complicates the application of pharmacogenetics within the country, as the one-drug-fits-all strategy is even less likely to be relevant. Additionally the reference metabolism of most drugs has been designed according to the less diverse Caucasian populations and is thus unlikely to comprehensively apply to the diverse African populations, which constitute the majority of the country, further emphasising a need for studies such as this one to be performed.

5.3.2 Frequency Comparison of CYP2C19 Metaboliser Classes

When comparing the metaboliser classes of the Xhosa to the Caucasian, Asian and CMA populations, the Asians show the highest frequency of PMs, the Xhosa show the highest frequency of IMs and the Caucasians show the highest frequency of UMs. The CMA population appears to form intermediate frequencies of the metaboliser groups between the populations (refer to Figure 5.9), suggestive once again of the high level of admixture observed in this population. The difference in frequencies of metaboliser classes observed between the Xhosa and CMA populations serves once again as a

reminder that the different populations residing in South Africa may need to be considered independently. Furthermore, it should be noted that the metaboliser classes reported in these two examined South African populations do not take any novel/uncharacterised variants into account, therefore after phenotypic validation of these uncharacterised variants, the frequencies of metaboliser classes may change.

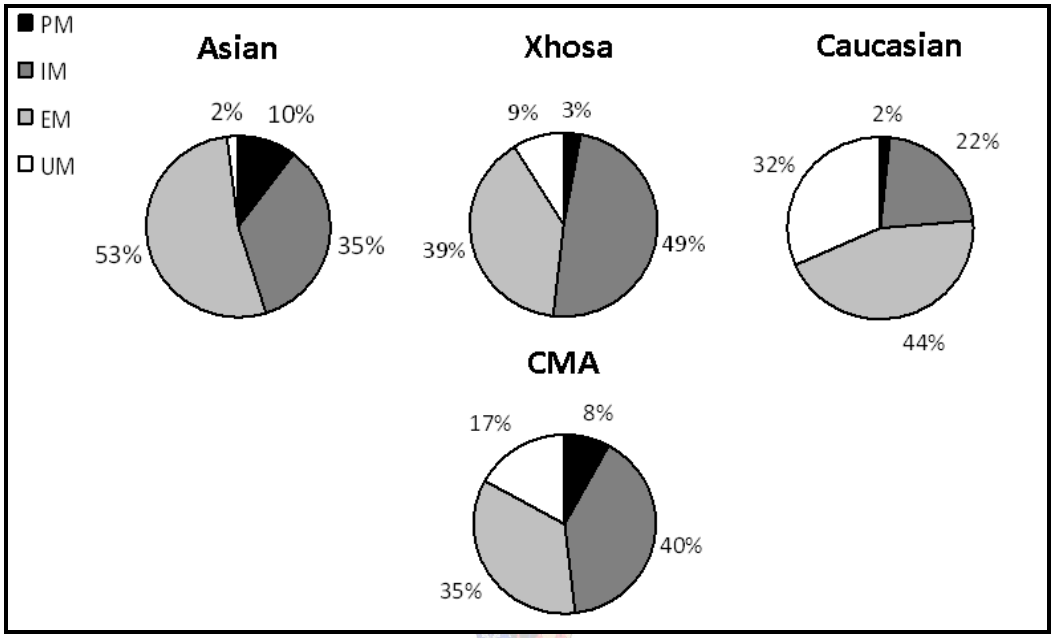


Figure 5.9: The frequencies of metaboliser classes observed in the Xhosa, Caucasian, Asian and CMA populations.
(Ragia *et al.* 2009, Chen *et al.* 2008, Da Silva *et al.* 2009)

It is important, to note that the pattern of metaboliser classes in different populations may influence the frequency of treatment failure or ADRs observed in the respective populations. For instance, in the Caucasian population, where a large frequency of UM individuals are observed, patients may experience a greater percentage of certain treatment failure. Conversely, in the Asian population, where a large percentage of PMs are detected, certain ADRs may be a more frequent occurrence. The same concept may be applied to the inducibility of CYP2C19 in the case of co-administration of drugs, as it has been reported that UMs show the greatest inducibility, while PMs exhibit the lowest inducibility (Caraco *et al.* 1995, 1996; Desta *et al.* 2002). Thus when considering the treatment of different populations, it may be advisable to bear in mind that certain populations may be more sensitive to the effects of concomitant medication than others.

When focussing on the frequency of metaboliser classes observed in the Xhosa population, there are a number of interesting factors. Firstly, the high percentage of IMs in this population is of significance when considering the high prevalence of HIV/AIDS in South Africa, with 11.6% of the

population reported to be infected with HIV/AIDS (http://www.unaids.org/en/Regions_Countries/Countries/south_africa.asp). This has potential implications for the treatment of these individuals, as it has been reported that in other genes with pharmacogenetic relevance (e.g. *CYP2D6*), IM individuals may shift towards PM metabolisers due to the additive effect that intermediate metabolism may have on the already compromised systems of these individuals (O'Neil *et al.* 2000). This may contribute towards the higher frequency of ADRs reported in individuals infected with HIV/AIDS (Mehta *et al.* 2007). As IM individuals are usually associated with a normal metabolism, this preconception may require re-evaluation in the context of HIV/AIDS in South Africa.

Furthermore, this study along with the data obtained for the CMA population in our laboratory, is the first to our knowledge to report a class of *CYP2C19* UMs in a South African population. The fact that the *CYP2C19* UM group has not been previously reported in South African populations is probably due to the lack of *CYP2C19*17* genotyping studies being performed. The presence of *CYP2C19* UM groups in South African populations may have greater implications than mere treatment failure, as it has been suggested that with regards to antidepressants, which are in part metabolised by CYP2C19 (Kirchheiner 2001; Herrlin *et al.* 2003; Wang *et al.* 2001), treatment failure may increase the risk of suicide. This may occur in UM individuals as they are more likely to experience treatment failure, which may in turn result in a greater likelihood of suicide, as has been previously reported for *CYP2D6* UMs (Zackrisson *et al.* 2009). Interestingly, a study comparing the rates of standardized mortality rates (SMR) of suicide in individuals of different ethnicity, found that the SMRs in Asian individuals were significantly lower (Bhui and McKenzie 2008). This is interesting as the Asian group showed the lowest frequency of UMs (refer to Figure 5.9). Therefore, it may be of value to use the metaboliser group data to guide the way in which the treatment plans of different populations are considered.

5.4 Other Mechanisms of Control and Areas of Interest Within and Around *CYP2C19*

5.4.1 Sequence Conservation

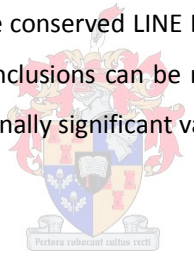
When considering areas of particular value to CYP2C19 expression and functioning, the level of sequence conservation, among other things, should be considered. Taking a closer look at the variants detected within the 5'-upstream regions, no variants were found in areas that have been proven to influence CYP2C19 expression (Gerbal-Chaloin *et al.* 2002) (refer to Figure 2.6). Interestingly, although only two individuals were re-sequenced for the area that has been reported to have been deleted without a significant effect on the expression of CYP2C19 (Arefayene *et al.* 2003), more than one third of the variants detected in the 5'-upstream region in this study, were located within this region, further validating the likely lack of functional significance of this area. Similarly, placing focus on the coding regions of the gene, it is interesting to note that in this

population, as well as the populations studied by Blaisdell *et el*. (2002) and Matimba *et al.* (2009), no variants were identified in exon 6. When consulting the Human *CYP2C19* Allele Nomenclature website, it was noticed that the lowest number of non-synonymous variants were detected in this exon. Interestingly, amino acids in this exon were found to be involved in the putative CYP2C19 active site (Lewis *et al.* 1998) and when the exonic sequences of the *Homo sapiens CYP2C19* gene were compared to the *Pan troglodytes* using BLAST (http://www.ncbi.nlm.nih.gov/), the sequences of exon 6 differed by only 1 bp. The lack of variation within functionally valuable areas as opposed to a relatively high level of variation within areas of no known functional value, within and between species, offers a method by which to highlight areas of possible importance to CYP2C19.

Using these principles of functional conservation, multi-species sequence comparisons were utilised to prioritise areas for the detection of 5'-upstream regulatory regions in which the occurrence of variants could have functional consequences for CYP2C19 expression. All the areas identified in the *Homo sapiens* by this analysis, fell between 6 405 bp and 7 164 bp upstream from the translational start site. After closer inspection, it was determined that all the identified areas of conservation were consistently located within repetitive elements, specifically the L1 class of LINEs, across all species. Subsequent consultation of the literature reported that LINE L1s have originated in all mammals from a common ancestor (Burton *et al.* 1986); therefore the similarity observed across the species may merely be an indication of a common ancestral sequence from which significant divergence has not yet taken place. It may be possible that these regions have remained conserved due to the fact that the selective pressure against these areas is not strong enough to cause a rapid accumulation of mutations and therefore a large divergence of these specific sequences between the different species (Monroe 2009).

This being said, it has been reported that LINE L1s may have a functional role to play in gene expression. It has been proposed that after LINE families undergo the retrotransposition phase, they will eventually gather mutations and become non-functional (Adey *et al.* 1994). It may, however, transpire that although these LINEs become inactive as a result of the accumulation of mutations, they may retain their 5'-upstream regulatory regions. This has been experimentally validated for the LINE L1 present in the 5'-upstream region of the apolipoprotein (a) (apo(a)) gene, which has been reported to increase the expression of apo(a) tenfold due to internal transcription factor binding sites present in the 5' region of the LINE (Yang *et al.* 1998). Similarly it has been reported that in avian species, the CR1 LINE, which is comparable to the mammalian LINE L1, was conserved as a result of functional properties. This conservation may be due to an enhancer region present in the LINE or due to the highly stable secondary structure created by the inverted LINE repeat clusters, which may affect the expression of nearby genes (Smith and Burgoyone 2001). Lastly it could be

important to bear in mind that it has been suggested that CpG islands may be responsible for interfering with the expression of inserted LINE sequences (Hata and Sakaki 1997), however, as no CpG island were detected in the 5'-upstream region of *CYP2C19*, this should not be the case for the LINEs observed in this study. It should also be noted that none of the conserved regions identified in the *Homo sapiens* were located in the 5' region of the LINE, but were found in the 3' end and open reading frame (ORF) 2. This does not however rule out the possible functional relevance of these areas, as the conserved regions, along with the other identified LINE L1s in this region, may be involved in creating a stable secondary structure and thereby affecting the expression levels of CYP2C19. Additionally new transcription factor binding sites could be created in these areas, as a result of the accumulation of mutations, which are responsible for controlling the expression of the downstream *CYP2C19* gene. After analysing the conserved regions, putative transcription factor binding sites for the four transcription factors controlling the expression of liver specific genes (Cereghini 1996) and acting as key regulators of CYP expression (Gonzales and Lee 1996; Rodriguez-Antona *et al.* 2002), namely C/EBP, HNF-1, HNF-3 and HNF-4, were all identified. This may signify an additional regulation system for CYP2C19 and thus a requirement for conservational pressure to be exerted on these regions. Although these conserved LINE L1 areas do provide an interesting avenue for future research, before any direct conclusions can be made with regards to the functionality of these regions and the detection of functionally significant variants, further experimental studies need to be performed.



It is also valuable to consider that although the high level of sequence conservation observed between the *CYP2C19* gene and other *Homo sapiens* and *Pan troglodytes CYP2C* genes, prevented the identification of individual areas of conservation, this obstacle may be used as a tool in future studies to identify regions that are specific to a particular gene's functioning. Therefore, in the future it may be beneficial to focus on the differences observed between these areas rather than the similarities. As each gene has its own specific function and level of expression, perhaps the subtle differences observed, may offer some valuable clues to the individual control and functioning of each gene. This is of special application to the *CYP2C* genes which exhibit high levels of sequence similarity, but differ substantially with regards to the level of expression of each gene (Goldstein *et al.* 1994).

5.4.2 CpG Island Analysis

Although CpG island analysis did reveal putative islands within the region containing the *Homo sapiens CYP2C* genes, the identified islands could not verify the different levels of expression observed between these four genes. The only correlation that could be made between the level of expression and the CpG islands identified, was that for both *CYP2C19* and *CYP2C18*, which reportedly

have the lowest *CYP2C* expression, putative CpG islands occurring between the start and stop codon were identified. Furthermore, analysis by Ingelman-Sunberg *et al.* (2007), utilising Methyl Primer Express, v1.0 (Applied Biosystems™, Foster City, California, USA) with a minimal CpG island length of 200 bp as opposed to 500 bp, identified an additional CpG island in intron 1 of *CYP2C19*. Although these identified islands do not occur within the 5'-upstream regions and are therefore unlikely to affect the availability of transcription factor binding sites, it has been suggested that methylated CpG islands present in other regions, such as intronic regions, may influence the conformation of the chromatin and thereby affect the expression levels of the associated gene (Rountree *et al.* 2001).

Four putative CpG islands were identified within and around the *CYP2C19* gene, including the one predicted by Ingelman-Sundberg *et al.* (2007), the presence of which could have implications for future pharmacogenetic applications. As epigenetics, and more specifically methylation of CpG islands, offers a link between environmental influences and gene expression, it would be of value to consider the effects that external factors have on the methylation status, and thus the expression of genes involved with the metabolism of drugs. It has been reported that diet (Cooney *et al.* 2002), smoking (Anttila *et al.* 2003) and alcohol consumption (Biermann *et al.* 2009) influence the level of methylation observed in certain CpG islands. Therefore if any of the CpG islands identified within and around *CYP2C19* are experimentally validated, the effect of these environmental factors on the validated islands may need to be considered in combination with *CYP2C19* genotype when determining the optimal drug dosage to deliver. Unfortunately, in practice, the elucidation of *CYP2C19* methylation status is not an easy procedure, as a liver biopsy will be required to determine the methylation status in the organ where metabolism by CYP2C19 occurs. The aim of pharmacogenetics is, among other things, to decrease the requirement for hospital procedures, for both patient comfort and economic reasons. An extra clinical procedure entailing sedation, medical expertise, equipment and possible complications (Beckmann *et al.* 2009) does not, therefore, serve this purpose. A possible alternative to liver biopsies would be to accumulate samples for liver biobanks, which may be used to perform the necessary methylation studies.

Methylation, however, is not the only avenue of epigenetic research relevant to pharmacogenetics and CYP2C19. Another mechanism that has received recent attention is the regulation of genes through miRNA directed mechanisms, whereby miRNAs recognises the 3'UTR of the mRNA of interest, thereafter degrading the mRNA (Lim *et al.* 2005). Although, to date only *CYP1B1* has been shown to be degraded in this manner (Tsuchiya *et al.* 2006), possibly as a result of the large 3'UTR, the expression levels of other CYP genes may also be controlled in a similar manner. Since the *CYP2C19* 3'UTR, is only 348 bp (http://www.ensembl.org/index.html), in comparison to the 3 119 bp of the *CYP1B1* 3'UTR, this gene does not appear to be one of the most likely candidates for this

mechanism of regulation. All the factors mentioned in this section, serve to highlight the endless complex mechanisms that require consideration in order to accurately determine the level and functioning of *CYP2C19*, emphasizing a need for future investigations.

CHAPTER 6:

CONCLUSIONS AND FUTURE PERSPECTIVES

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

The aim of this study was to detect the presence and frequency of common variation within the *CYP2C19* gene in the Xhosa population and to elucidate the effect of any novel/uncharacterised variants identified by this analysis, utilising both *in silico* and *in vitro* analyses. Furthermore, after the prioritisation of the SNPs identified within the Xhosa population, the aim was to compare the genetic variation observed in this population to other populations, both local and foreign, and to use this data to devise appropriate genotyping panels that could be implemented in other populations. Lastly, various *in silico* techniques were used to elucidate possible areas of control for future studies.

This project successfully detected 30 variants occurring within the *CYP2C19* gene in the South African Xhosa population. Utilising *in silico* analyses, a possible functional role for the -2030T variant, which was found to occur in perfect LD with the *CYP2C19*15* variants, was suggested. Although dual luciferase reporter assays did not convincingly validate the predicted role of the -2030T variant, the dual reporter luciferase assays did show a significant decrease in the fold induction observed with regards to the construct containing the rs7902257 variant. This variant was designated *CYP2C19*27* and was found at a frequency of 0.33 in the Xhosa population, which is significantly higher than the frequencies previously reported in other populations (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=7902257).

In addition to these variants, a novel non-synonymous V374I variant was designated *CYP2C19*28* and was detected in the Xhosa population at a frequency of 0.01. Furthermore, the allele-defining variants for *CYP2C19*2*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17* were detected at a frequency of 0.21, 0.09, 0.09 and 0.10, respectively. All these variants, were successfully genotyped and detected in the CMA population (Da Silva *et al.*, 2009), proving the successful application of this genotyping protocol in other South African populations. It was, however, shown that the genetic profiles of different populations varied from one another, suggesting that the genetic profile and thus pharmacogenetic treatment of one population may not be directly applied to another population. This was demonstrated by the differences observed in the genetic profiles of the Caucasian, Asian, CMA and African American populations in comparison to the Xhosa population (refer to Figures 5.4, 5.7 and 5.9). The genetic profiles of more closely related populations, such as the Xhosa and Venda populations, may however be used to guide the pharmacogenetic treatment of each other.

Lastly, the *in silico* analysis used to highlight additional areas of interest, prioritised four regions within the *CYP2C19* 5'-upstream region that have been conserved, all of which occur within LINE L1s,

and which may be conserved as a result of some functional significance. Furthermore, CpG island analysis revealed three additional putative islands that may be involved in the epigenetic control of CYP2C19 and should therefore be considered for future studies.

Although this project has led to the much needed characterisation of *CYP2C19* within the South African Xhosa population, the data generated requires phenotypic validation. Additionally, as only 15 healthy Xhosa individuals were sequenced, rare variants present in the Xhosa population may have been missed. Therefore, it is of vital importance that the results obtained from this study, are validated through the procedures depicted in Figure 2.2 that have not been carried out by this study. This should initially be implemented through genotype-phenotype studies, to identify phenotypic outlier individuals, which may subsequently be genotyped for the variants detected in this study. All individuals displaying outlier phenotypes in the absence of any of the previously detected variants will undergo subsequent sequence analysis in order to identify additional variants that may affect the functionality of the CYP2C19 enzyme, thereby accounting for the previously overlooked rare variants. Once the genotype-phenotype correlations have been made, the study should be replicated in an independent Xhosa population and the variants shown to affect phenotype should be implemented into clinical trials. It is important that this study moves above and beyond the academic scope and that the clinical utility of the results are confirmed. Only then, can the data generated start to benefit the South African community.

The high level of sequence variation present in the Xhosa population emphasizes the unique challenges that South Africa faces with regards to the application of pharmacogenetics in the Xhosa and other South African populations. Genotyping tests such as the Roche AmpliChip test (http://www.amplichip.us/) are already available for *CYP2C19*. However, although this test does constitute the first FDA approved *CYP2D6* and *CYP2C19* pharmacogenetics test, this test is not able to detect the *CYP2C19*9*, *CYP2C19*15*, *CYP2C19*17* or novel/uncharacterised variants. Thus, in the context of the genetically diverse African populations, other genotyping protocols will most likely be required. Even though sequence analysis is not the most cost-effective method at present, sequencing technologies are rapidly advancing, becoming a cheaper, easier and more informative alternative to most other genotyping methods. To give an idea of how quickly sequencing technology has advanced; in 2003 the sequencing of the human genome took 13 years and cost $2.7 billion, while in 2007 sequencing the human genome took two months and cost <$1 million (Louie 2007). The use of sequence analysis would be of great value to the *CYP* genes where variation is common and the occurrence of adjacent variants such as *CYP2C19*2* and *CYP2C19*10* may lead to misclassification, where RFLP analysis is utilised. Additionally, African populations stand to benefit from sequence analysis due to the prevalence of rare variants which would otherwise be missed, but

could contribute to non-optimal treatment outcomes. Thus, this project has served as a reminder that the pharmacogenetic tests which are designed according to a template Caucasian population may not be applicable to the diverse African populations. Furthermore, it is essential that in the future comprehensive, innovative and cost effective genotyping techniques for the successful pharmacogenetic characterisation of African populations are devised.

Although the information generated by this project is anticipated to play a role in decreasing ADRs and treatment failure in cases where *CYP2C19* genotype has been convincingly associated with therapeutic response to drugs such as tamoxifen (Schroth *et al.* 2007) and clopidogrel (Shuldiner *et al.* 2009), it is important to note that the data obtained for *CYP2C19* comprises only one of the many factors that need to be considered in pharmacogenetic applications. In order for pharmacogenetics to be successfully applied, a comprehensive database for each unique population is required, including as much information on as many genes as possible. Among the unique South African populations for which pharmacogenetic profiles remain to be elucidated, the Afrikaner, Indian populations as well as representatives from the other African populations falling in clusters 2 and 3 depicted in Figure 2.12 may present interesting avenues for future studies.

With regards to drug metabolism in the Xhosa population, studies have already been performed to identify the common variation present in this population for *CYP2D6* (Wright *et al.* 2009, manuscript submitted to *Prog Neuropsychopharmacol Biol Psychiatry*), *CYP3A4* (Warnich *et al.* 2009, manuscript in preparation) and *CYP2C19*. Additionally, data is also available for *CYP2C9* from a study performed on a black South African population (Mitchell *et al.* 2008) and data for both *CYP2C19* (Da Silva *et al.* 2009) and *CYP2D6* (Gaedigk and Coetsee 2008) has been obtained for the CMA population. When referring to the most important *CYP* genes as reported by Ingelman-Sundberg *et al.* (2007), the four genes in the above-mentioned South African studies were reported to constitute more than half of the visits made to the *CYP* websites, hinting at their value both to academic and industry related fields. Thus, when considering the metabolism of therapeutic agents in South African populations, the data generated from all these genes should be considered in combination with one another, especially in the case where two or more enzymes are collectively involved in the metabolism of the therapeutic agent under examination. In addition to the combined examination of the genes involved in the metabolism themselves, it may be important, in future, to merge this data with information obtained with regards to transcription factors and other areas that are involved in the regulation of the genes, such as the putative areas identified by this study.

It must be noted that the genotype of an individual is only a partial component with regards to the effective metabolism of a drug. There are many other external factors that may influence

metabolism such as smoking, diet, concomitant drugs, physiological or disease status, age and demographic factors (Sotanieui *et al.* 1997; Kashuba *et al.* 1998; Ingelman-Sundberg *et al.* 1999; Arranz and de Leon 2007), which may act on the genes themselves through epigenetic mechanisms or which may, by virtue of their interaction with the ingested drugs, influence the metabolism outcome. Thus in order for realistic dosage recommendations to be made, the effect of all these factors in combination with one another, needs to be taken into account. As this may become extremely complicated, appropriate algorithms will need to be developed. Such algorithms have already been generated with regards to warfarin dosage recommendation (Sasaki *et al.* 2009). Although, at present we are in the initial phases of pharmacogenetics in South Africa, and elucidation of variation in combination with clinical validation remains a priority, the end goal should remain in sight. Thus, appropriate algorithms should be considered during the process of this data generation.

Once the effect of all factors involved has been determined and genotyping panels, dosage recommendations and clinical procedures have been elucidated, it remains essential that the disciplines work together to reduce the occurrence of ADRs and treatment failure. By involving those that are directly affected by the occurrence of ADRs and treatment failure, such as the patients and clinicians, the therapeutic aspects which affect the patients to the greatest extent and thus require urgent addressing, may be identified. When referring to the research performed by Okezie and Olufunmilayo (2008), it appears that even when reporting forms are available to document the occurrence of ADRs, most people are unaware of their existence. Thus, in order to incorporate the community into the research and therefore ensure that the end goal of pharmacogenetics is obtained, comprehensive information needs to be made available. This strategy could improve patient compliance by involving the patients and by creating awareness that by optimising treatment plans, serious ADRs, treatment failure and even minor side effects such as weight gain can be eliminated. This may be especially useful in the context of HIV/AIDS where patients receiving ART are more likely to experience ADRs (Mehta *et al.* 2007) and a careful treatment balance needs to be determined. It is essential that the benefits of treatment outweigh the harm caused by serious ADRs and in order to do this all individuals involved, from the lab to the patient, need to be consulted.

There is much potential for the steadily growing field of personalised medicine, the market of which is presently estimated at $232 billion and is expected to grow to $500 billion by 2015 (http://www.fiercebiotech.com/press-releases/). The data generated by this study is, to our knowledge, the most comprehensive for *CYP2C19* in a South African population and validated the hypothesis that the genetic profile of *CYP2C19* in the Xhosa population differs from that of other previously studied populations. Therefore, this study should be used to build on pharmacogenetic research in South Africa. Future studies should focus on validating the results obtained in this study

in independent populations and designing tests that are applicable to South African populations. The data generated from these studies should be utilized by means of a holistic approach, where algorithms are incorporated to combine all factors. By implementing all of these factors, the eventual aim of this research would be to aid in the alleviation of the treatment associated side-effects of drugs that are metabolised by CYP2C19, in the context of South Africa.

REFERENCES

Adey NB, Schichman SA, Graham DK, Peterson SN, Edgell MH, Hutchison CA: Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol. Biol. Evol*. 11, 778-789 (1994).

Allison M: Is personalized medicine finally arriving? *Nat. Biotechnol.* 26, 509-517 (2008).

Andersson T, Holmberg J, Rohss K, Walan A: Pharmacokinetics and effect on caffeine metabolism of the proton pump inhibitors, omeprazole, lansoprazole, and pantoprazole. *Br. J. Clin. Pharmacol.* 45, 369–375 (1998).

Andersson T, Flockhart DA, Goldstein DB *et al.*: Drug-metabolizing enzymes: evidence for clinical utility of pharmacogenomic tests. *Clin. Pharmacol. Ther.* 78, 559–581 (2005).

Anttila S, Hakkola J, Tuominen P *et al.*: Methylation of cytochrome P4501A1 promoter in the lung is associated with tobacco smoking. *Cancer Res*. 63, 8623–8628 (2003).

Aoyama N, Tanigawara Y, Kita T *et al.*: Sufficient effect of 1-week omeprazole and amoxicillin dual treatment for Helicobacter pylori eradication in cytochrome P450 2C19 poor metabolizers. *J. Gastroenterol.* 34, 80–83 (1999).

Aparicio S, Chapman J, Stupka E *et al.*: Whole-genome shotgun assembly and analysis of the genome of Fugu rubribes. *Science* 297, 1301-1310 (2002).

Arefayene M, Skaar TC, Zhao X *et al.*: Sequence diversity and functional characterization of the 5'-regulatory region of human CYP2C19. *Pharmacogenetics* 13, 199-206 (2003).

Arranz MJ, de Leon J: Pharmacogenetics and pharmacogenomics of schizophrenia: a review of the last decade of research. *Mol. Psych.* 12, 707-741 (2007).

Assenat E, Gerbal-chaloin S, Maurel P, Vilarem MJ, Pascussi JM: Is nuclear factor kappa-B the missing link between inflammation, cancer and alteration in hepatic drug metabolism in patients with cancer? *Eur. J. Cancer.* 42, 785-792 (2006).

Backstrom M, Mjorndal T, Dahlqvist R, Nordkvist-Olsson T: Attitudes to reporting adverse drug reactions in northern Sweden. *Clin. Pharmacol.* 56, 729–732 (2000).

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

Bathum L, Skejlbo E, Mutabingwa TK, Madsen H, Horder M, Brosen K: Phenotypes and genotypes for CYP2D6 and CYP2C19 in a black Tanzanian population. *Br. J. Clin. Pharmacol.* 48, 395-401 (1999).

Bauer M, Whybrow PC, Angst J, Versiani M, Moller HJ: World federation of societies of biological psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders, part 1: acute and continuation treatment of major depressive disorder. *World J. Biol. Psychiatry* 3, 5-43 (2002).

Beckmann MG, Bahr MJ, Hadem J *et al.*: Clinical relevance of transjugular liver biopsy in comparison with percutaneous and laparoscopic liver biopsy. *Gastroenterol. Res. Pract.* 1-7 (2009).

Belton KJ, Lewis SC, Payne S, Rawlins MD, Wood SM: Attitudinal survey of adverse drug reaction reporting by medical practitioners in the United Kingdom. *Clin. Pharmacol.* 39, 223–226 (1995).

Bertilsson L, Henthorn TK, Sanz E, Tybring G, Säwe J, Villén T: Importance of genetic factor in the regulation of diazepam metabolism: relationship to S-mephenytoin, but not debrisoquin, hydroxylation phenotype. *Clin. Pharmacol. Ther.* 45, 348-355 (1989).

Bertz RJ, Granneman GR: Use of in vitro data to estimate the likelihood of metabolic pharmacokinetic interactions. *Clin. Pharmacol.* 32, 210-258 (1997).

Bhat R, Weaver JA, Sterling KM, Bresnick E: Nuclear transcription factor Oct-1 binds to the 5' upstream region of CYP1A1 and negatively regulates its expression. *Int. J. Biochem. Cell Biol.* 28, 217-227 (1996).

Bhui KS, McKenzie K: Rates and risk factors by ethnic group for suicides within a year of contact with mental health services in England and Wales. *Psychiatr. Serv.* 59, 414-420 (2008).

Biermann T, Reulbach U, Lenz B *et al.*: N-methyl-D-aspartate 2b receptor subtype (NR2B) promoter methylation in patients during alcohol withdrawal. *J. Neural. Transm.* 116, 615-622 (2009).

Bikandi J, San Millán R, Rementeria A, Garaizar J: In silico analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction. *Bioinformatics* 20, 798-799 (2004).

Boucher HW, Groll AH, Chiou CC, Walsh TJ: Newer systemic antifungal agents: pharmacokinetics, safety and efficacy. *Drugs* 64, 1997–2020 (2004).

Brosen K, de Morais SM, Meyer UA, Goldstein JA: A multifamily study on the relationship between CYP2C19 genotype and S-mephenytoin oxidation phenotype. *Pharmacogenetics* 5, 312-317 (1995).

Brown TA: Genomes 2. 2nd ed. United Kingdom (UK): BIOS Scientific Publishers; (2002).

Brown CS, Farmer RG, Soberman JE, Eichner SF: Pharmacokinetic factors in the adverse cardiovascular effects of antipsychotic drugs. *Clin. Pharmacokinet.* 43, 33-56 (2004).

Brunak S, Engelbrecht J, Knudsen S: Prediction of Human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220, 49-65 (1991).

Burton FH, Loeb DD, Voliva CF, Martin SL, Edgell MH, Hutchison CA: Conservation throughout Mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. *J. Mol. Biol.* 187, 291-304 (1986).

Calza L, Rosseti N, Biagetti C, Pocaterra D, Colangeli V, Manfredi R: Abacavir-induced reaction with fever and severe skin rash in a patient tested human leukocyte antigen-B*5701 negative. *Int. J. STD AIDS*. 20, 276-277 (2009).

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

Caraco Y, Tateishi T, Wood AJ: Interethnic difference in omeprazole's inhibition of diazepam metabolism. *Clin. Pharmacol. Ther.* 58, 62-72 (1995).

Caraco Y, Wilkinson GR, Wood AJ: Differences between white subjects and Chinese subjects in the in vivo inhibition of cytochrome P450s, 2C19, 2D6, and 3A by omeprazole. *Clin. Pharmacol. Ther.* 60, 396-404 (1996).

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

Cereghini S: Liver-enriched transcription factors and hepatocyte differentiation. *FASEB J.* 10, 267-282 (1996).

Chang M, Dahl ML, Tybring G, Gotharson E, Bertilsson L: Use of omeprazole as a probe drug for CYP2C19 phenotype in Swedish Caucasians: comparison with S-mephenytoin hydroxylation phenotype and CYP2C19 genotype. *Pharmcogenetics* 5, 358-363 (1995).

Chen Y, Ferguson SS, Negishi M, Goldstein A: Identification of constitutive androstane receptor and glucocorticoid receptor binding sites in the CYP2C19 promoter. *Mol. Pharmacol*. 64, 316-324 (2003).

Chen L, Qin S, Xie J *et al.*: Genetic polymorphism analysis of CYP2C19 in Chinese Han populations from different geographic areas of mainland China. *Pharmacogenomics* 9, 691-702 (2008).

Chen Y, Goldstein JA: The transcriptional regulation of the human CYP2C genes. *Curr. Drug Metab.* 10, 567-578 (2009).

Cooney CA, Dave AA, Wolff GL: Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *J. Nutr.* 132, 2393S–2400S (2002).

Cox E, Mager D, Weisbart E: Geographic variation trends in prescription use:2000 to 2006. St Louis: Express Scripts (2008).

Cross J, Lee H, Westelinck A, Nelson J, Grudzinskas C, Peck C: Post marketing drug dosage changes of 499 FDA-approved new molecular entities, 1980-1999. *Pharmacoepidemiol. Drug. Saf.* 11, 439-446 (2002).

Dandara C, Masimirembwa CM, Magimba A *et al.*: Genetic polymorphism of CYP2D6 and CYP2C19 in East- and Southern African populations including psychiatric patients. *Eur. J. Clin. Pharmacol.* 57, 11-17 (2001).

Da Silva D, Drögemöller BI, Wright GEB, Warnich L: Assessment of the genetic variation in CYP2C19 in the South African Mixed Ancestry population. Hons Thesis. Stellenbosch University (2009).

Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M: Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *PLoS ONE*. 4, e4439 (2009).

Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR: Genetic epidemiology and public health: hope, hype, and future prospects. *The Lancet.* 366, 1484–1498 (2005).

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

de Leon J, Susce MT, Johnson M *et al.*: DNA microarray technology in the clinical environment: the AmpliChip CYP450 test for CYP2D6 and CYP2C19 genotyping. *CNS Spectr*. 14, 19-34 (2009).

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

de Morais SM, Wilkinson GR, Blaisdell J, Meyer UA, Nakamura K, Goldstein JA: Identification of a new genetic defect responsible for the polymorphism of (S)-mephenytoin metabolism in Japanese**.** *Mol. Pharmacol.* 46, 594-598 (1994b).

de Smith AJ, Walters RG, Coin LJ *et al.*: Small deletion variants have stable breakpoints commonly associated with alu elements. *PLoS ONE* 3, e3104 (2008).

Desta Z, Zhao X, Shin J-G, Flockhart DA: Clinical significance of the cytochrome P450 2C19 genetic polymorphism. *Clin. Pharmacokinet.* 41, 913-958 (2002).

Dogan RI, Getoor L, Wilbur WJ Mount SM: SplicePort: an interactive splice-site analysis tool. *Nucleic Acids Res.* W285-291 (2007).

Dossing M, Pilsgaard H, Rasmussen B, Poulssen HE: Time course of phenobarbital and cimetidine mediated changes in hepatic drug metabolism. *Eur. J. Clin. Pharmacol.* 25, 215-222 (1983).

Duan S, Zhang W, Cox NJ, Dolan ME: FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation*.3, 139-141 (2008).

Ehret C: Bantu expansions: re-envisioning a central problem of early African history. *Int. J. Afr. Hist. Stud.* 34, 5-41 (2001).

Eichelbaum M, Ingelman-Sundberg M, Evans WE: Pharmacogenomics and individualized drug therapy. *Annu. Rev. Med.* 57, 119−137 (2006).

Eland IA, Elton KJ, van Grootheest AC, Meiners AP, Rawlins MD, Stricker BH: Attitudinal survey of voluntary repotting of adverse drug reactions. *Clin. Pharmacol.* 48, 623–627 (1999).

Ellis KJ, Stouffer GA, McLeod HL, Lee CR: Clopidogrel pharmacogenomics and risk of inadequate platelet inhibition: US FDA recommendations. *Pharmacogenomics* 10, 1799-1817 (2009).

Entsuah AR, Huang H, Thase ME: Response and remission rates in different subpopulations with major depressive disorder administrated venlaxine, selective serotonin reuptake inhibitors, or placebo. *J. Clin. Psychiatry* 62, 869-877 (2001).

Ferguson RJ, de Morais SM, Benhamou S *et al.* : A new genetic defect in human CYP2C19: mutation of the initiation codon is responsible for poor metabolism of S-mephenytoin. *J. Pharmacol. Exp. Ther.* 284, 356-361 (1998).

Ferguson SS, LeCluyse EL, Negishi N, Goldstein JA: Regulation of human CYP2C9 by the constitutive androstane receptor: discovery of a new distal binding site. *Mol. Pharmacol.* 62, 737-746 (2002).

Ferguson SS, Chen Y, LeCluyse EL, Negishi N, Goldstein JA: Human CYP2C8 is transcriptionally regulated by the nuclear receptors constitutive androstane receptor, pregnane X receptor, glucocorticoid receptor and hepatic nuclear factor 4α. *Mol. Pharmacol.* 68, 747-757 (2005).

Fiala-Beer E, Lee AC, Murray M: Regulation of the rat CYP2A4 gene promoter by c-Jun and octamer binding protein-1. *Int. J. Biochem. Cell. Biol.* 39, 1235-1247 (2007).

Flockhart DA, McMillin GA: A pharmacogenomics approach to using warfarin. ARUP laboratories, salt lake city, UT. http://arup.cnpg.com/med/webinar/lokhartmcmillin/20060829. 1-11 (2006).

Ford GA, Wood SM, Daly AK: CYP2D6 and CYP2C19 genotypes of patients with terodiline cardiotoxicity identified through the yellow card system. *Br. J. Clin. Pharmacol.* 50, 77-80 (2000).

Frueh FW, Amur S, Mummaneni P *et al.*: Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy* 28, 992-998 (2008).

Furuta T, Ohashi K, Kamata T *et al.*: Effect of genetic differences in omeprazole metabolism on cure rates for Helicobacter pylori infection and peptic ulcer. *Ann. Intern. Med.* 129, 1027-1030 (1998).

Furuta T, Ohashi K, Kosuge K *et al.*: CYP2C19 genotype status and effect of omeprazole on intragastric pH in humans. *Clin. Pharmacol. Ther.* 65, 552–561 (1999).

Furuta T, Shirai N, Takashima M *et al.*: Effect of genotypic differences in CYP2C19 on cure rates for Helicobacter pylori infection by triple therapy with a proton pump inhibitor, amoxicillin, and clarithromycin. *Clin. Pharmacol. Ther.* 69, 158–168 (2001).

Gaedigk A, Coetsee C: The CYP2D6 gene locus in South African Coloureds: unique allele distributions, novel alleles and gene arrangements. *Eur. J. Clin. Pharmacol.* 64, 465-475 (2008).

Gaikovitch EA, Cascorbi I, Mrozikiewiez PM *et al.*: Polymorphisms of drug metabolising enzymes CYP2C9, CYP2C19, CYP2D6, CYP1A1, NAT2 and of P-glycoprotein in a Russian population. *Eur. J. Clin. Pharmacol.* 59, 303-312 (2003).

Garcia-Barceló M, Chow LY, Kum Chiu HF *et al.*: Frequencies of defective CYP2C19 alleles in a Hong Kong Chinese population: detection of the rare allele CYP2C19*4. *Clin. Chem.* 45, 2273-2274 (1999).
Gardiner SJ, Begg EJ: Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. *Pharmacol. Rev.* 58, 521–590 (2006).

Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF: Evidence or archaic Asian ancestry on the human X chromosome. *Mol. Biol. Evol.* 22, 189-192 (2004).

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

Gerbal-Chaloin S, Daujat M, Pascussi J-M, Pichard-Garcia L, Vilarem M-J, Maurel P: Transcriptional regulation of CYP2C9 gene: role of glucocorticoid receptor and constitutive androstane receptor. *J. Biol. Chem.* 277, 209-217 (2002).

Gidal BE, Zupanc ML: Potential pharmacokinetic interaction between felbamate and phenobarbital. *Ann. Pharmacother.* 28, 455-458 (1994).

Goldstein JA, Faletto MB, Romkes-Sparks M *et al.:* Evidence that CYP2C19 is the major (S)-mephenytoin 4'-hydroxylase in humans. *Biochemistry 33, 1743-1752* (1994).

Goldstein JA, de Morais SMF: Biochemistry and molecular biology of the human CYP2C subfamily. *Pharmacogenetics* 4, 285-299 (1994).

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

Gomez A, Ingelman-Sundberg M: Pharmacoepigenetics: its role in interindividual differences in drug response. *Clin. Pharmacol. Ther.* 85, 426-430 (2009).

Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA: Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* 24, 757–768 (2007).

Gonzalez F, Lee Y: Constitutive expression of hepatic cytochrome P450 genes. *FASEB J.* 10, 1112-1117 (1996).

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

Guengerich FP: Cytochrome P450s and other enzymes in drug metabolism and toxicity. *AAPS J*. 8, E101-E111 (2006).

Haas DW, Smeaton LM, Shafer RW *et al.*: Pharmacogenetics of long-term responses to antiretroviral regimens containing efavirenz and/or nelfinavir: an adult aids clinical trials group study. *J. Infect. Dis.* 192, 1931−1942 (2005).

Hall TA: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.* 41, 95-98 (1999).

Halling J, Peterson MS, Damkier P *et al.*: Polymorphism of CYP2D6, CYP2C19, CYP2C9 and CYP2C8 in the Faroese population. *Eur. J. Clin. Pharmacol.* 61, 491-497 (2005).

Hamdy SI, Hiratsuka M, Narahara K *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.* 53, 596-603 (2002).

Harding RM, McVean G: A structural ancestral population for the evolution of modern humans. *Curr. Opin. Genet. Dev.* 14, 667-674 (2004).

Hardison RC: Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16, 369–372 (2000).

Hasford J, Goettler M, Munter KH, Muller-Oerlinghausen B: Physicians' knowledge and attitudes regarding the spontaneous reporting system for adverse drug reactions. *Epidemiology* 55, 945–950 (2002).

Hastings PJ, Ira G, Lupski JR: A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.*. 5, e1000327 (2009).

Hata K, Sakaki Y: Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene.* 189, 227-234 (1997).

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: Splice site prediction in Arabidopsis thaliana DNA by combining local and global sequence information. *Nucleic Acids Res.* 24, 3439-3452 (1996).

Helsby NA, Ward SA, Edwards G, Howells RE, Breckenridge AM: The pharmacokinetics and activation of proguanil in man: consequences of variability in drug metabolism. *Br. J. Clin. Pharmacol.* 30, 593-598 (1990).

Herrlin K, Yasui-Furokori N, Tybring G, Widen J, Gustafsson LL, Bertilsson L: Metabolism of citalopram enantiomers in CYP2C19/CYP2D6 phenotyped panels of healthy Swedes. *Br. J. Clin. Pharmacol.* 56, 415–421 (2003).

Hoskins JM, Shenfield GM, Gross AS: Concordance between proguanil phenotype and CYP2C19 genotype in Chinese. *Eur. J. Clin. Pharmacol.* 59, 611-614 (2003).

Hughes HB, Biehl JP, Jones AP, Schmidt LH: Metabolism of isoniazid in man as related to the occurrence of peripheral neuritis. *Am. Rev. Tuberc*. 70, 266 (1954).

Hulot JS, Bura A, Villard E *et al.*: Cytochrome P450 2C19 loss-of-function polymorphism is a major determinant of clopidogrel responsiveness in healthy subjects. *Blood* 108, 2244−2247 (2006).

Ibeanu GC, Blaisdell J, Ferguson RJ *et al.*: A novel transversion in the intron 5 donor splice junction of CYP2C19 and a sequence polymorphism in exon 3 contribute to the poor metabolizer phenotype for the anticonvulsant drug S-mephenytoin. *J. Pharmacol. Exp. Ther.* 290, 635-640 (1999).

Ingelman-Sundberg M, Oscarson M, McLellan RA: Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *TiPS* 20, 342-349 (1999).

Ingelman-Sundberg M: Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, the present and future. *Drug Metab. Dispos.* 25, 193-200 (2004).

Ingelman-Sundberg M: The human genome project and novel aspects of cytochrome P450 research. *Toxicol. Appl. Pharmacol.* 207, S52-S56 (2005).

Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C: Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol. Ther.* 116, 496–526 (2007).

Inomata S, Nagashima A, Itagaki F *et al.*: CYP2C19 genotype affects diazepam pharmacokinetics and emergence from general anesthesia. *Clin. Pharmacol. Ther.* 78, 647–655 (2005).

Janha RE, Sisay-Joof F, Hamid-Adiamoh M *et al.*: Effects of genetic variation at the CYP2C19/CYP2C9 locus on pharmacokinetics of chlorcycloguanil in adult Gambians. *Pharmacogenomics* 10, 1423-1431 (2009).

Justenhoven C, Hamann U, Pierl CB *et al.*: CYP2C19*17 is associated with decreased breast cancer risk. *Brest Cancer Res. Treat.* 115, 391-396 (2009).

Kaneko A, Bergqvist Y, Taleo G, Kobayakawa T, Ishizaki T, Björkman A: Proguanil disposition and toxicity in malaria patients from Vanuatu with high frequencies of CYP2C19 mutations. *Pharmacogenetics* 9, 317-326 (1999).

Kashuba ADM, Bertino JS, Rocci ML, Kulaway RW, Beck DJ, Nafziger AN: Quantification of 3-month intraindividual variations and influence of sex and menstral cycle phases on CYP3A4 activity as measured by phenotyping with intravenous midazolam. *Clin. Pharmacol. Ther.* 64, 269-277 (1998).

Kawashima S, Kobayashi K, Takama K *et al.*: Involvement of the hepatocyte nuclear factor 4α in the different expression level between CYP2C9 and CYP2C19 in the human liver. *Drug Metab. Dispos.* 34, 1012-1018 (2006).

Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*31, 3576-3579 (2003).

Kerb R, Fux R, Mörike K *et al.*: Pharmacogenetics of antimalarial drugs: effect on metabolism and transport. *Lancet Infect. Dis.* 9, 760-774 (2009).

Khaliq Y, Gallicano K, Seguin I *et al.*: Single and multiple dose pharmacokinetics of nelfinavir and CYP2C19 activity in human immunodeficiency virus-infected patients with chronic liver disease. *Br. J. Clin. Pharmacol.* 50, 108-115 (2000).

Kilicarslan T, Haining RL, Rettie AE, Busto U, Tyndale RF, Sellers EM: Flunitrazepam metabolism by cytochrome P450s 2C19 and 3A4. *Drug Metab. Dispos.* 29, 460-465 (2001).

Kirchheiner J, Brosen K, Dahl ML *et al.*: CYP2D6 and CYP2C19 genotype-based dose recommendations for antidepressants: a first step towards subpopulation-specific dosages. *Acta. Psychiatr. Scand.* 104, 173-192 (2001).

Kirchheiner J, Nickchen K, Wong M-L, Licinio J, Roots I, Brockmöller J: Pharmacogenetics of antidepressants and antipsychotics: the contribution of allelic variations to the phenotype of drug response. *Mol. Psych.* 9, 442-473 (2004).

Kirchheiner J, Fuhr U, Brockmoller J: Pharmacogenetics-based therapeutic recommendations-ready for clinical practice? *Nat. Rev. Drug. Discov.* 4, 639-647 (2005).

Klug WS, Cummings MR: Concepts of genetics. 7th ed. New Jersey: Pearson Education Inc (2003).

Knodell RG, Dubey RK, Wilkinson GR, Guengerich FP: Oxidative metabolism of hexobarbital in human liver: relationship to polymorphic S-mephenytoin 4-hydroxylation. *J. Pharmacol. Exp. Ther.* 245, 845-849 (1988).

Kubota T, Chiba K, Ishizaki T: Genotyping of S-mephenytoin 4'hydroxylation in an extended Japanese population. *Clin. Pharmacol. Ther.* 60, 661-666 (1996).

Kudla G, Murray AW, Tollervey D, Plotkin JB: Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324, 255-258 (2009).

Kupfer A, Preisig R: Pharmacogenetics of mephenytoin. A new drug hydroxylation polymorphism in man. *Eur. J. Clin. Pharmacol.* 26, 753-759 (1984).

Lahn BT, Ebenstein L: Let's celebrate human genetic diversity. *Nature* 461, 726-728 (2009).

Lane AB, Soodyall H, Arndt S *et al.*: Genetic substructure in South African Bantu-speakers: evidence from autosomal DNA and Y-chromosome studies. *Am. J. Phys. Anthropol.* 119, 175–185 (2002).

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

Lee SH, Wang X-L, DeJong J: Functional interactions between atypical NF-κB site from the rat CYP2B1 promoter and the transcriptional repressor RBP-Jκ/CBF1. *Nuc. Acid Res.* 28, 2091-2098 (2000).

Levy RH: Cytochrome P450 isoenzymes and antiepileptic drug interactions. *Epilepsia* 36, 8S-S13 (1995).

Lewis DF, Dickins M, Weaver RJ, Eddershaw PJ, Goldfarb PS, Tarbit MH: Molecular modeling of human CYP2C subfamily enzymes CYP2C9 and CYP2C19: rationalization of substrate specificity and site-directed mutagenesis experiments in the CYP2C subfamily. *Xenobiotica* 28, 235-268 (1998).

Lewis DF: 57 varieties: the human cytochromes P450. *Pharmacogenomics* 5, 305–318 (2004).

Lim LP, Lau NC, Garrett-Engele P *et al.*: Microarray analysis shows that some microRNAs down regulate large numbers of target mRNAs. *Nature* 433, 769−773 (2005).

Llerena A, Berecz R, Dorado P, Gonzalez AP, Penas-LLedo EM, De la Rubia A: CYP2C9 gene and susceptibility to major depressive disorder. *Pharmacogenomics J.* 3, 300-302 (2003).

Loots G, Ovcharenko I, Pachter L, Dubchak I, Rubin E: rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. *Genome. Res.* 12, 832-839 (2002).

Louie AS: Jumpstarting the transformation to true personalized medicine: potential catalysts and keystones. *Health Industry Insights* #HI208059 (2007).

Mabadeje AFB, Akintola AA, Ashorobi RA: The value and effects of implementing an essential drugs list in the Lagos University Teaching Hospital. *Clin. Pharmacol. Ther.* 50, 121–124 (1991).

Mann HB, Whitney DR: On a test of whether one of two random variables is stochastically larger than the other. *Ann. Appl. Probab.* 18, 50–60 (1947).

Manolopoulos VG: Pharmacogenomics and adverse drug reactions in diagnostic and clinical practice. *Clin. Chem. Lab. Med.* 45, 801–814 (2007).

Masimirembwa CM, Hasler JA: Genetic polymorphisms of drug metabolizing enzymes in African populations: implications for the use of neuroleptics and antidepressants. *Brain Res. Bull.* 44, 561-571 (1997).

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

McElroy JP, Nelson MR, Caillier SJ, Oksenberg JR: Copy number variation in African Americans. *BMC Genet* 10, 15 (2009).

McKinnon RA, Evans AM: Cytochrome P450: 2. Pharmacogenetics, *Aus. J. Hosp. Pharm.* 30, 102–105 (2000).

Meadows M: DA approves heart drug for black patients. *FDA Consum.* 39, 8-9 (2005).

Meaney MJ, Szyf M: Maternal care as a model for experience-dependant chromatin plasticity. *Trends Neurosci.* 28, 456-463 (2005).

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

Miller SA, Dykes DD, Polesky HF: A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16, 1215 (1988).

Miller M: Tools for population genetic analyses (TFPGA) 1.3: a Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by the author (1997).

Mitchell C: Evaluation of CYP2C9 and VKORC1 gene variants that may result in warfarin dosage sensitivity and poor pregnancy outcomes. MSc thesis, Wits University (2008).

Miura J, Obua C, Abbo C, Kaneko S, Tateishi T: Cytochrome P450 2C19 genetic polymorphisms in Ugandans. *Eur. J. Clin. Pharmacol.* 65, 319-320 (2009).

Monroe D: Genetics. Genomic clues to DNA treasure sometimes lead nowhere. *Science* 325, 142-143 (2009).

Munos B: Can open-source R&D reinvigorate drug research? *Nat. Rev. Drug Discov.* 5, 723-729 (2006).

Murray CJ, Lopez AD: Global mortality, disability, and the contribution of risk factors: global burden of disease study. *Lancet* 349, 1436-1442 (1997).

Nakamoto K, Kidd JR, Jenison RD *et al.*: Genotyping and haplotyping of CYP2C19 functional alleles on thin-film biosensor chips. *Pharmacogenet. Genomics* 17, 103-114 (2007).

Nebert DW, Russell DW: Clinical importance of the cytochromes P450. *Lancet* 360, 1155-1162 (2002).

Nelson DR, Koymans L, Kamataki T *et al.*: P450 superfamily: update on new sequence, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* 6, 1-42 (1996).

Nelson MH, Dolder CR: Lapatinib: A novel dual tyrosine kinase inhibitor with activity in solid tumors. *Ann. Pharmacother.* 40, 261-269 (2006).

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

Okezie EO, Olufunmilayo F: Adverse drug reactions reporting by physicians in Ibadan, Nigeria. *Pharmacoepidemiol. Drug Saf.* 17, 517–522 (2008).

Olsen H, Koppang E, Alvan G, Mørland J: Carisoprodol elimination in humans. *Ther. Drug Monit.* 16, 337-340 (1994).

O'Neil WB, Gilfix BM, Markoglou N, Di Girolamo A, Tsoukas CM, Wainer IW: Genotype and phenotype of cytochrome P450 2D6 in human immunodeficiency virus-positive patients and patients with acquired immunodeficiency syndrome. *Eur. J. Clin. Pharmacol.* 56, 231-240 (2000).

Ortiz de Montellano PR: Cytochrome P450: structure, mechanism and Biochemistry. 3rd ed. New York: Plenum Publishers (2005).

Oscarson M: Pharmacogenetics of drug metabolising enzymes: importance for personalised medicine. *Clin. Chem. Lab. Med.* 41, 573-580 (2003).

Patel KJ, Kedia MS, Bajpai D, Mehta SS, Kshirsagar NA, Gogtay NJ: Evaluation of the prevalence and economic burden of adverse drug reactions presenting to the medical emergency department of a tertiary referral centre: a prospective study. *BMC Clinical Pharmacology* 7, 8 (2007).

Patterson N, Petersen DC, van der Ross RE *et al.*: Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* doi:10.1093/hmg/ddp505 (2009).

Peires J: The dead will arise: Nongqawuse and the great Xhosa cattle-killing movement of 1856-7. Johannesburg: Raven Press. (1989).

Persson I, Aklillu E, Rodrigues F, Bertilsson L, Ingelman-Sundberg M: S-mephenytoin hydroxylation phenotype and CYP2C19 genotype among Ethiopians. *Pharmacogenetics* 6, 521-6 (1996).

Pheasant M, Mattick J: Raising the estimate of functional human sequences. *Genome Res.* 17, 1245-1253 (2007).

Pirmohamed M, James S, Meakin S, Green C *et al.*: Adverse drug reactions as cause of admission to hospital. *Br. Med. Jr*. 329, 15-19 (2004).

Plagnol V, Wall JD: Possible ancestral structure in human populations. *PLoS Genet.* 2:e105 (2006).

Prabhakar S, Poulin F, Shoukry M *et al.*: Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.* 16, 855-862 (2006).

Prior TI, Chue PS, Tibbo P, Baker GB: Drug metabolism and atypical antipsychotics. *Eur. Neuropsychopharmacology* 9, 301-309 (1999).

Quandt K, Frech K, Karas H, Wingender E, Werner T: MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* 23, 4878-4884 (1995).

Ragia G, Arvanitidis KI, Tavridou A, Manolopoulos VG: Need for reassessment of reported CYP2C19 allele frequencies in various populations in view of CYP2C19*17 discovery: the case of Greece. *Pharmacogenomics* 10, 43-49 (2009).

Ramensky V, Bork P, Sunyaev S: Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900 (2002).

Rasmussen H: Misclassification of allele CYP2C19*10 as CYP2C19*2 by a commonly used PCR-RFLP procedure. *Genet. Test.* 12, 57-58 (2008).

Redon R, Ishikawa S, Fitch KR *et al.*: Global variation in copy number in the human genome. *Nature* 444, 444–454 (2006).

Reed FA, Tishkoff SA: African human diversity, origins and migrations. *Curr. Opin. Genet. Dev.* 16, 597-605 (2006).

Ridley, M: Evolution. Oxford: Blackwell Science Ltd (2004).

Rodriguez-Antona MT, Donato A, Boobis RJ *et al.*: Cytochrome P450 expression in human hepatocytes and hepatoma cell lines: molecular mechanisms that determine lower expression in cultured cells. *Xenobiotica* 32, 505-520 (2002).

Roh HK, Dahl ML, Tybring G, Yamada, Cha YN, Bertilsson L: CYP2C19 genotype and phenotype determined my omeprazole in a Korean population. *Pharmacogenetics* 6, 547-551 (1996).

Roses AD: Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nat. Rev.* 5, 645–656 (2004).

Rountree MR, Bachman KE, Herman JG, Baylin SB: DNA methylation, chromatin inheritance, and cancer. *Oncogene* 20, 3156−3165 (2001).

Rozen S, Skaletsky HJ: Primer3 on the www for general users and for biologist programmers. *Methods Mol. Biol.* 132, 365–386 (2000).

Sagar M, Seensalu R, Tybring G, Dahl ML, Bertilsson L: CYP2C19 genotype and phenotype determined with omeprazole in patients with acid-related disorders with and without Helicobacter pylori infection. *Scand. J. Gastroenterol.* 33, 1034-1038 (1998).

Sagar M, Janczewska I, Ljungdahl A, Bertilsson L, Seensalu R: Effect of CYP2C19 polymorphism on serum levels of vitamin B12 in patients on long-term omeprazole treatment. *Aliment Pharmacol. Ther.* 13, 453-458 (1999).

Sagar M, Tybring G, Dahl M-L, Bertilsson L, Seensalu R: Effects of omeprazole on intragastric pH and plasma gastrin are dependent on the CYP2C19 polymorphism. *Gastroenterology* 119, 670–676 (2000).

Sapone A, Vaira D, Trespidi S *et al.*: The clinical role of cytochrome P450 genotypes in Helicobacter pylori management. *Am. J. Gastroenterol.* 98, 1010–1015 (2003).

Sasaki T, Tabuchi H, Higuchi S, Leiri I: Warfarin-dosing algorithm based on a population pharmacokinetic/pharmacodynamic model combined with Bayesian forecasting. *Pharmacogenomics*. 10, 1257-1266 (2009).

Schroth W, Antoniadou L, Fritz P *et al.*: Breast cancer treatment outcome with adjuvant tamoxifen relative to patient *CYP2D6* and *CYP2C19* genotypes. *J. CLin. Oncol.* 25, 5187-5193 (2007).

Schweizer E, Rynn M, Mandos LA, Demartinis N, Garcia-Espana F, Rickels K: The antidepressant effect of sertraline is not enhanced by dose titration: results from an outpatient clinical trial. *Int. Clin. Psychopharmacol.* 16, 137−143 (2001).

Seedat S, Emsley RA, Stein DJ: Land of promise: challenges and opportunities for research in South Africa. *Mol. Psychiatry* 9, 891–892 (2004).

Shapiro SS, Wilk MB: An analysis of variance test for normality (complete samples). *Biometrika.* 52. 591–611 (1965).

Shimada T, Yamazaki H, Mimura M, Nui Y, Guengerih FP: Interindividual variation in human liver cytochrome P450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians. *Pharmacol. Exp. Ther.* 270, 414-423 (1994).

Shirai N, Furuta T, Moriyama Y *et al.*: Effects of CYP2C19 genotypic differences in the metabolism of omeprazole and rabeprazole on intragastric pH. *Aliment. Pharmacol. Ther.* 15, 1929–1937 (2001).

Shuldiner A, O'Connell JR, Bliden KP *et al.*: Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA.* 302, 849-858 (2009).

Sim SC, Risinger C, Dahl ML *et al.*: A common novel CYP2C19 gene variant causes ultrarapid drug metabolism relevant for the drug response to proton pump inhibitors and antidepressants. *Clin. Pharmacol. Ther.* 79, 103-113 (2006).

Singh OP, Bali P, Hemingway J, Subbarao SK, Dash AP, Adak T: PCR-based methods for the detection of L1014 kdr mutation in Anopheles culicifacies sensu lato. *Malar J.* 14, 154 (2009).

Sistonen J, Fusellib S, Paloa JU, Chauhanc N, Padhc H, Sajantilaa A: Pharmacogenetic variation at CYP2C9, CYP2C19, and CYP2D6 at global and microgeographic scales. *Pharmacogenet. Genomics* 19, 170–179 (2009).

Smith LM, Burgoyne LA: Species identity: conserved inverted LINE repeat clusters (ILRC) in the vertebrate genome as indicators of population boundaries. *Gene* 271, 273-283 (2001).

Sotanieui EA, Arranto AJ, Pelkonen O, Pasanen M: Age and cytochrome P450-linked drug metabolism in humans: an analysis of 226 subjects with equal histopathological conditions. *Clin. Pharmacol. Ther.* 61, 331-339 (1997).

Steemers FJ, Gunderson KL: Whole genome genotyping technologies on the BeadArray platform. *Biotechnol. J.* 2, 41–49 (2007).

Steimer W, Muller B, Leucht S, Kissling W: Pharmacogenetics: a new diagnostic tool in the management of antidepressive drug therapy. *Clin. Chim. Acta.* 308, 33-41 (2001).

Steimer W, Zopf K, von Amelunxen S, Pfeiffer H, Bachofer J, Popp J: Amitriptyline or not, that is the question: pharmacogenetic testing of CYP2D6 and CYP2C19 identifies patients with low or high risk for side effects in amitriptyline therapy. *Clin. Chem.* 51, 376−385 (2005).

Stranger BE, Forrest MS, Dunning M *et al.*: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848−853 (2007).

Sugimoto K, Uno T, Yamazaki H, Tateishi T: Limited frequency of the CYP2C19*17 allele and its minor role in a Japanese population. *Br. J. Clin. Pharmacol.* 65, 437-439 (2008).

Takada K, Arefayene M, Desta Z *et al.*: Cytochrome P450 pharmacogenetics as a predictor of toxicity and clinical response to pulse cyclophosphamide in lupus nephritis. *Arthritis Rheum.* 50, 2202−2210 (2004).

Takai D, Jones PA: Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U S A.* 99, 3740-3745. (2002).

Tanigawara Y, Aoyama N, Kita T, Shirakawa K, Komada F, Kasuga M, Okumura K: CYP2C19 genotype-related efficacy of omeprazole for the treatment of infection called by Helicobacter pylori. *Clin. Pharmacol. Ther.* 66, 528–534 (1999).

Tenesa A, Navarro P, Hayes BJ *et al.*: Recent human effective population size estimated from linkage disequilibrium. *Genome Res*. 17, 520-526 (2007).

Thase ME: Effectiveness of antidepressants: comparative remission rates. *J. Clin. Psychiatry* 64, 3-7 (2003).

Thum T, Haverich A, Borlak J: Cellular dedifferentiation of endothelium is linked to activation and silencing of certain nuclear transcription factors: implications for endothelial dysfunction and vascular biology. *Faseb J.* 14, 740-751 (2000).

Tishkoff SA, Dietzsch E, Speed W *et al.*: Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271, 1380-1387 (1996).

Tishkoff SA, Verrelli BC: Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* 4, 293-340 (2003).

Tishkoff SA, Reed FA, Friedlaender FR *et al.*: The genetic structure and history of Africans and African Americans. *Science* 324, 1035-1044 (2009).

Tsuchiya Y, Nakajima M, Takagi S, Taniya T, Yokoi T: MicroRNA regulates the expression of human cytochrome P450 1B1. *Cancer Res.* 66, 9090–9098 (2006).

Vogel F: Moderne probleme der humangentik. *Ergebnisse der Inneren Medizin* 12, 52-125 (1959).

Wan J, Xia H, He N, Lu YQ, Zhou HH: The elimination of diazepam in Chinese subjects is dependent on the mephenytoin oxidation phenotype. *Br. J. Clin. Pharmacol.* 42, 471-474 (1996).

Wang JH, Liu ZQ, Wang W *et al.*: Pharmacokinetics of sertraline in relation to genetic polymorphism of CYP2C19. *Clin. Pharmacol. Ther.* 70, 42-47 (2001).

Wanwimolruk S, Bhawan S, Coville PF, Charlcroft SCW: Genetic polymorphism of debrisoquine (*CYP2D6)* and proquanil (*CYP2C19*) in South Pacific Polynesian populations. *Eur. J. Clin. Pharmacol.* 54, 431-435 (1998).

Warnich L, Niehaus DJH, Plummer M *et al.*: Common allelic variants of the CYP3A4 gene in the Khoisan, Xhosa and Mixed Ancestry populations from South Africa. Manuscript in preparation.

Wester K, Jonnson AK, Sigset O, Druid H, Hagg S: Incidence of fatal adverse drug reactions: a population based study. *Br. J. Clin. Pharmacol.* 65, 573-579 (2008).

Wiffen P, Gill M, Edwards J, Moore A: Adverse drug reactions in hospital patients. A systematic review of the prospective and retrospective studies. *Bondolier Extra.* 1-16 (2002).

Wilkins PB, Wrighton SA, Schuetz EG, Molona DT, Guzelian PS: Identification of glucocortioid inducible cytochrome P450 in the intestinal mucosa of rats and man. *J. Clin. Inves.* 80, 1029-1036 (1987).

Willard HF, Ginsburg GS: Genomic and personalized medicine, 1st ed. London: Elsevier; vol 1, Chp 27 and 29 (2009).

Wright GEB, Niehaus DJH, Drögemöller BI, Koen L, Warnich L: Elucidation of CYP2D6 genetic diversity in a unique African population: implications for Xhosa schizophrenia patients. Manuscript submitted to *Prog Neuropsychopharmacol Biol Psychiatry* (2009).

Xie HG, Kim RB, Stein CM, Wilkinson GR, Wood AJJ: Genetic polymorphism of (S)-mephenytoin 4'hydroxylation in populations of African descent. *Br. J. Clin. Pharmacol.* 48, 402–408 (1999).

Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R: Apolipoprotein(a) Gene Enhancer Resides within a LINE Element. *J. Biol. Chem.* 273, 891–897 (1998).

Yokono A, Morita S, Someya T, Hirokane G, Okawa M, Shimoda K: The effect of CYP2C19 and CYP2D6 genotypes on the metabolism of clomipramine in Japanese psychiatric patients. *J. Clin. Psychopharmacol.* 21, 549−555 (2001).

Yotova V, Lefebvre JF, Kohany O *et al.*: Tracing genetic history of modern humans using X-chromosome lineage. *Hum. Genet.* 122, 431-443 (2007).

Yu KS, Yim DS, Cho JY *et al.*: Effect of omeprazole on the pharmacokinetics of moclobemide according to the genetic polymorphism of CYP2C19. *Clin. Pharmacol. Ther.* 69, 266−273 (2001).

Zackrisson AL, Lindblom B, Ahlner J: High frequency of occurrence of CYP2D6 gene duplication/multiduplication indicating ultrarapid metabolism among suicide cases. *Clin. Pharmacol. Ther.* 2009 Nov 11. [Epub ahead of print].

Zuker M: Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406-3415 (2003).

Electronic sources

$232 Billion Personalized Medicine Market to Grow 11 Percent Annually, says PricewaterhouseCoopers
http://www.fiercebiotech.com/press-releases/232-billion-personalized-medicine-market-grow-11-percent-annually-says-pricewaterhous#ixzz0ZIGGU0OK
Accessed December 2009

1000 Genomes Project:
http://www.1000genomes.org/page.php?page=home
Accessed October 2009

AmpliChip
http://www.amplichip.us/
Accessed February 2009

ClustalW
http://www.ebi.ac.uk/Tools/clustalw2/index.html
Accessed March 2008

CpG island searcher:
http://www.cpgislands.com/
Accessed February 2009

CpG Plot:
http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html
Accessed February 2009

CYP2C19 allele nomenclature:
http://www.cypalleles.ki.se/cyp2c19.htm
Accessed August 2009

Database of Genomic Variants
http://projects.tcag.ca/variation
Accessed October 2009

Ensembl Genome Browser
http://www.ensembl.org/index.html
Accessed November 2007

Ethnologue: Languages of the world
http://www.ethnologue.com
Accessed March 2008

FDA. Guidance for Industry Pharmacogenomic Data Submissions (2005):
http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126957.pdf
Accessed June 2008

Global Health Atlas
http://www.who.int/GlobalAtlas/predefinedReports/EFS2008/index.asp?strSelectedCountry=ZA
Accessed August 2009

HapMap3:
http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap3r2_B36/
Accessed April 2009

Human Genome Project
www.genome.gov/HGP
Accessed June 2009

IDTSci OligoAnalyser v3.1.
http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/
Accessed June 2008

National Human Genome Research Institute
www.genome.gov/HGP
Accessed February 2008

NCBI dbSNP:
http://www.ncbi.nlm.nih.gov/snp/
Accessed January 2008

Protocol: TaqMan® Copy Number Assays
https://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_062368.pdf

Promega: Complete Current Vector List
http://www.promega.com/vectors/allvectors.htm

RestrictionMapper
http://www.restrictionmapper.org
Accessed 2008

SNPs3D
http://www.snps3d.org/
Accessed September 2009

Statistics South Africa (Census 2001):
http://www.statssa.gov.za/census01/HTML/default.asp
Accessed August 2009

The Malaria Research Programme of the Medical Research Council, South Africa
http://www.malaria.org.za/
Accessed October 2009

UNAIDS. South Africa:
http://www.unaids.org/en/Regions_Countries/Countries/south_africa.asp
Accessed August 2009

WebCutter 2.0
http://rna.lundberg.gu.se/cutter2/
Accessed 2008

APPENDIX 1: HUMAN *CYP2C19* ALLELE NOMENCLATURE

(http://www.cypalleles.ki.se/cyp2c19.htm)

Allele	***Protein***	***Nucleotide changes***		***Trivial name***	***Effect***	***Enzyme activity***		***References***
		cDNA	***Gene**** **To be considered CYP2C19*1A, the genomic reference sequence must be changed to 99C and 80161A.*			***In_vivo***	***In_vitro***	
*CYP2C19*1A*	CYP2C19.1A	None	None		None	Normal	Normal	Romkes *et al.*, 1991
*CYP2C19*1B*	CYP2C19.1B	99C>T; 991A>G	99C>T; 80161A>G		I331V	Normal		Richardson *et al.*, 1997
*CYP2C19*1C*	CYP2C19.1B	991A>G	80161A>G		I331V	Normal		Blaisdell *et al.*, 2002
*CYP2C19*2A*		99C>T; **681G>A**; 990C>T; 991A>G	99C>T; **19154G>A**; 80160C>T; 80161A>G	m1; m1A	**Splicing defect**; I331V	None		de Morais *et al.*, 1994a
*CYP2C19*2B*		99C>T; 276G>C; **681G>A**; 990C>T; 991A>G	99C>T; 12460G>C; **19154G>A**; 80160C>T; 80161A>G	m1B	E92D; **splicing defect**; I331V	None		Ibeanu *et al.*, 1998b
*CYP2C19*2C (also called CYP2C19*21)*		99C>T; 481G>C; **681G>A**; 990C>T; 991A>G	-98T>C; 99C>T; 12122G>A; 12662A>G; 12834G>C; **19154G>A**; 19520A>G; 57740C>G; 79936T>A; 80160C>T; 80161A>G		A161P, **splicing defect**, I331V			Fukushima-Uesaka *et al.*, 2005
*CYP2C19*2D*		99C>T; **681G>A**; 990C>T; 991A>G; 1213G>A	-98T>C; 99C>T; 12662A>G; **19154G>A**; 57740C>G; 80160C>T; 80161A>G; 87275G>A		**splicing defect**, E405K			Lee *et al.*, 2009
*CYP2C19*3A*		**636G>A**; 991A>G; 1251A>C	**17948G>A**; 80161A>G; 87313A>C	m2	**W212X**; I331V	None		de Morais *et al.*, 1994b
*CYP2C19*3B (also called CYP2C19*20)*		**636G>A**; 991A>G; 1078G>A; 1251A>C	-889T>G; 12013T>G; 12122G>A; 12306G>A; 13166T>C; **17948G>A**; 18911A>G; 80161A>G; 80248G>A; 87313A>C		**W212X**; D360N; I331V			Fukushima-Uesaka *et al.*, 2005
*CYP2C19*4*		**1A>G**; 99C>T, 991A>G	**1A>G**; 99C>T; 80161A>G	m3	**GTG initiation codon**; I331V	None		Ferguson *et al.*, 1998
*CYP2C19*5A*	CYP2C19.5A	**1297C>T**	**90033C>T**	m4	**R433W**	None	None	Xiao *et al.*, 1997 Ibeanu *et al.*, 1998a

*CYP2C19*5B*	CYP2C19.5B	99C>T; 991A>G; **1297C>T**	99C>T; 80161A>G; **90033C>T**		I331V; **R433W**	None		Ibeanu *et al.*, 1998a
*CYP2C19*6*	CYP2C19.6	99C>T; **395G>A**; 991A>G	99C>T; **12748G>A**; 80161A>G	m5	**R132Q**; I331V	None	None	Ibeanu *et al.*, 1998b
*CYP2C19*7*			**19294T>A**		**Splicing defect**	None		Ibeanu *et al.*, 1999
*CYP2C19*8*	CYP2C19.8	**358T>C**	**12711T>C**		**W120R**	None	Decr	Ibeanu *et al.*, 1999
*CYP2C19*9*	CYP2C19.9	99C>T; **431G>A**; 991A>G	99C>T; **12784G>A**; 80161A>G		**R144H**; I331V		(Decr)	Blaisdell *et al.*, 2002
*CYP2C19*10*	CYP2C19.10	99C>T; **680C>T**; 991A>G	99C>T; **19153C>T**; 80161A>G		**P227L**; I331V		Decr	Blaisdell *et al.*, 2002
*CYP2C19*11*	CYP2C19.11	99C>T; 449G>A; 991A>G	99C>T; 12802G>A; 80161A>G		R150H; I331V			Blaisdell *et al.*, 2002
*CYP2C19*12*	CYP2C19.12	99C>T; 991A>G; **1473A>C**	99C>T; 80161A>G; **90209A>C**		I331V; **X491C**; 26 extra aa		Unstable	Blaisdell *et al.*, 2002
*CYP2C19*13*	CYP2C19.13	991A>G; 1228C>T	80161A>G; 87290C>T		I331V; R410C			Blaisdell *et al.*, 2002
*CYP2C19*14*	CYP2C19.14	50T>C; 99C>T; 991A>G	50T>C; 99C>T; 80161A>G		L17P; I331V			Blaisdell *et al.*, 2002
*CYP2C19*15*	CYP2C19.15	55A>C; 991A>G	55A>C; 80161A>G		I19L; I331V			Blaisdell *et al.*, 2002
*CYP2C19*16*	CYP2C19.16	1324C>T Existence of the *CYP2C19*2* polymorphism 681G>A on the same allele can not be excluded	90060C>T		R442C			Morita *et al.*, 2004
*CYP2C19*17*	CYP2C19.1	99C>T; 991A>G	-3402C>T; -1041A>G; **-806C>T**; 99C>T; 80161A>G		I331V	Incr.	Incr. transcr.	Sim *et al.*, 2006 Rudberg *et al.*, 2007
*CYP2C19*18*	CYP2C19.18	99C>T; 986G>A; 991A>G	99C>T; 80156G>A; 80161A>G; 87106T>C		R329H; I331V			Fukushima-Uesaka *et al.*, 2005
*CYP2C19*19*	CYP2C19.19	99C>T; 151A>G; 991A>G	99C>T; 151A>G; 80161A>G; 87106T>C		S51G; I331V			Fukushima-Uesaka *et al.*, 2005

*CYP2C19*20*	*See CYP2C19*3B*							
*CYP2C19*21*	*See CYP2C19*2C*							
*CYP2C19*22*	CYP2C19.22	557G>C; 991A>G	17869G>C; 80161A>G		R186P; I331V			Matimba *et al.*, 2009
*CYP2C19*23*	CYP2C19.23	99C>T; 271G>C; 991A>G	99C>T; 12455G>C; 80161A>G		G91R; I331V			Zhou *et al.*, 2009
*CYP2C19*24*	CYP2C19.24	99C>T; 991A>G; 1004G>A; 1197A>G	99C>T; 80161A>G; 80174G>A; 87259A>G		I331V; R335Q			Zhou *et al.*, 2009
*CYP2C19*25*	CYP2C19.25	99C>T; 991A>G; 1344C>G	99C>T; 80161A>G; 90080C>G		I331V; F448L			Zhou *et al.*, 2009
*CYP2C19*26*	CYP2C19.26	99C>T; 766G>A; 991A>G	99C>T; 19239G>A; 80161A>G		D256N; I331V			Lee *et al.*, 2009
Additional SNPs, where the haplotype has not yet been determined								
		221T>C			M74T			Solus *et al.*, 2004
		502T>C			F168L			Solus *et al.*, 2004
		636G>T/C/A			W212C/C/X			NCBI dbSNP
		337G>A			V113I			Matimba *et al.*, 2009
		905C>G			T302R			Chen *et al.*, 2008
		1180G>A			V394M			Lee *et al.*, 2009

Nucleotide variations that are underlined are nonsynonymous mutations.
Nucleotide variations in bold are the major SNPs/alterations responsible for the phenotype of the corresponding allele.

APPENDIX 2: CONSENT FORMS

PATIENT INFORMATION AND INFORMED CONSENT

GENETICS OF SCHIZOPHRENIA

(Where italics, please delete what is not applicable)

PURPOSE:

The Department of Psychiatry (University of Stellenbosch) and qualified researchers from other research institutions worldwide invites you/the participant to participate in this trial. Worldwide continuous discoveries are being made about different diseases or conditions due to research on the cells and molecules of the body. Therefore, in this trial we plan to investigate the role of genetics/inheritance in the outcome of first episode psychosis. In order to achieve this, genetic material (blood) will be collected to be analysed for certain defects and excess material will then be stored for future research.

STUDY PROCEDURE:

If you/the participant decides to participate, we will ask you/the participant questions to find out what symptoms you/the participant experience. Furthermore, recent findings indicate that some schizophrenia sufferers are born with unique physical characteristics. With the help of measurements and photographs of your face, hands and feet, we hope to identify such features. No photograph will be published that may reveal your identity. You/the participant will also be asked to complete a computer-based concentration and face recognition test. A urine sample to test for the presence of certain drugs in your/the participant's body will also be taken. You/the participant will also be requested to allow a qualified person to draw approximately 48 ml (3 tablespoons) of blood from your/the participant's arm after you have fasted for 12 hours. This blood will be used to test your blood sugar level, cholesterol level, prolactin level and to extract DNA for testing. After your/the participant's and your/the participant's immediate family's written informed consent is obtained, blood (also 48ml) may be requested from your/the participant's immediate family to investigate underlying familial tendencies in your/the participant's condition. These blood samples may be used to create a cell line. This is done by changing some of the blood cells so that they can grow indefinitely. The cell line is living tissue and it can be used to make more DNA or its products at any time in the future. The DNA or its products will then be taken from the cell line and saved for scientific analyses which will be performed now and possibly in the future."

We would like your/the participant's permission to contact your/the participant's relatives in order to get more information about any family history of mental illness. You /the participant can still participate in the study even if your/the particpant's relatives do not.

Personal information that could be used to identify you/the participant (such as name, contact information, etc) will not be given out. The data and DNA is likely to be made available to qualified scientists around the world to study this particular disorder.

Your/the participant's unidentified cell line and DNA will be maintained permanently, unless you/the participant request to

have it removed. If at any time in the future you/the participant wishes to have the DNA, cell lines or clinical data removed from the storage site, you/the participant may do so by contacting the researchers conducting this study (Prof Dana Niehaus at 021 - 938 9505).

The researchers who will have access to your/the participant's DNA include those who work with private and for profit companies who also study DNA collected in individuals with schizophrenia. These researchers may be interested in eventually developing commercial medical products by using the DNA from you/the participant and other participants. They may sell or patent discoveries based on this research and thus benefit financially. Please note that you/the participant or your /the participant's heirs will not receive any compensation if this occurs.

We do not expect to discover any information of direct benefit to your/the participant's condition, or treatment, during the next few years. If later on, diagnostic tests or new ways to treat the condition are discovered, this information will have to be obtained from properly licensed clinical labs, clinics, or your/the participant's physician, and will not be available from the research team.

RISKS:

You/the participant may feel some pain associated with having blood withdrawn from a vein. Discomfort, bruising and/or other bleeding at the site where the needle is inserted may also be experienced. Occasionally, some people experience fleeting dizziness or feel faint when their blood is drawn.

Some insurance companies may mistakenly assume that your/the participant's participation in this study is an indication of a higher risk of a genetic disease, and this could hurt access to health or other insurance. We will not share any information about you/the participant, or your/the participant's family, with an insurance company. However, if you/the participant discusses participation in this study with a doctor, and he or she records it in your/the participant's medical record, it is possible that an insurance company may access the information as part of a medical record review. It is the opinion of the investigators that this study is not genetic testing. It is aimed at developing such testing for the future, but cannot currently provide any meaningful information about participants. Against this background, this project should upon inquiry not be reported as genetic testing.

BENEFITS:

Other than your/the participant's current psychiatric condition being appropriately treated, there are no direct benefits to you/the participant at this stage. However, your/the participant's and other individuals' family and future generations may benefit if we can scientifically delineate these disorders, and thereby facilitate the rational approach to the clinical diagnosis and therapy of its manifestations, or locate the genes that lead to such disorders. That knowledge could then lead to the development of methods for prevention and forms of new treatments aimed at curing or alleviating these diseases.

CONFIDENTIALITY:

If you/the participant consents to participate in this study, your/the participant's identity will be kept confidential. Answers will not be shared with other family members or anyone else except for staff members involved in this study. All data will be

kept in locked file cabinets accessible only to the research staff. All research information obtained will not be associated with your/the participant's name; research staff will use only a coded number and/or your/the participant's initials. Blood samples will be safely stored and identified by code number and access will be limited to authorized scientific investigators. Copies of treatment records from hospitals or mental health professionals are kept in locked files and are reviewed by members of the research team only. Any publications/lectures/reports resulting from this study will not identify you/the participant by name.

VOLUNTARY PARTICIPATION:

Your/the participant's participation in this study is voluntary and you/the participant may refuse to participate or withdraw from the study at any time without any loss of benefits to which you/the participant are otherwise entitled. Some members of the team of investigators conducting this study may be responsible for your/the participant's clinical care, which will not be compromised if you/the participant refuses to participate.

RESEARCH QUESTIONS AND CONTACTS:

If you/the participant are interested in genetic counseling, information about where such counseling is given will be supplied. A new blood sample may be required if such counseling is attended. DNA information about a relative will be released only if the genetic counselor confirms that the relative in question is deceased or cannot be found and that the information is essential for clinical counseling.

The researchers will answer any questions you/the participant might have about the procedures described above, or about the results of the study. If you/the particpant have any questions, please call Dr Dana Niehaus (021) 938 9161.

The University of Stellenbosch Faculty of Health Sciences' Committee for Human Research has approved recruitment and participation of individuals for this study on the basis of: 1. Guidelines on Ethics for Medical Research of the SA Medical Research Council, 2. Declaration of Helsinki, 3. International Guidelines: Council for International Organisations for Medical Sciences (CIOMS), 4. Applicable RSA legislation.

I/the participant confirm that:

I have been given a copy of this consent form to keep.

The above information was explained in Afrikaans/English/Xhosa/Other _____, a language in which I am fluent. If needed, the translation and explanation was done by _____

I was afforded adequate time to pose questions and that these questions were fully answered.

I was not pressurized to participate.

I will not be paid for participation, but reimbursement of travel costs will be considered (where applicable).

I will not incur any additional costs through participation.

INFORMED CONSENT:

I /the participant have read the above patient information and hereby consent voluntarily to participate/allow the potential participant to participate in this study:

Print name: _____ Signature: _____

or right thumb print of participant/ representative of participant

Signed/Confirmed at _____ on _____20__

Signature of Witness_____

DECLARATION BY OR ON BEHALF OF INVESTIGATOR(S)

I _____ declare that I explained the information in this document to _____ and/or his/her representative _____ He/she was encouraged and afforded adequate time to ask me any questions. This conversation was conducted in _____ and no translator was used/ was translated into _____ by _____

Signed at _____ on _____20__

Signature of investigator/Representative of Investigator_____

Signature of Witness_____

DECLARATION OF TRANSLATOR

I _____ confirm that I translated the contents of this document from English to _____ and explained the contents of this document to the participant/participant's representative. I also translated questions posed by _____ as well as the answers given by _____ I conveyed a factually correct version of what was related to me.

Signed at _____ on _____20__

Signature of Translator_____

Signature of Witness _____

IMPORTANT MESSAGE TO PARTICIPANT/REPRESENTATIVE OF PARTICIPANT:

Dear participant/representative of participant

Thank you very much for your/the participant's participation in this study. Should, at any time during the study:

An emergency arise as a result of the research, or

You require any further information with regard to the study, kindly contact

_____at_____

_____

APPENDIX 3: SPECIFIED PROTOCOLS

3.1 Miller *et al.* 1988 gDNA Extraction Protocol

1. Shake the tube with blood well to mix the contents and transfer the contents (±10 ml) to a marked 50 ml polypropylene tube
2. Add ±30 ml cold Lysis Buffer and mix by inversion
3. Place the tube on ice for 15-30 min and mix by inversion every 5 min
4. Centrifuge the tubes for 10 min at 1 500 x g (4°C)
5. Carefully discard the supernatant and keep the pellet. Pat slightly dry on paper
6. Add 10 ml cold PBS to the pellet, mix and centrifuge again for 10 min at 1 500 x g (4°C)
7. Carefully discard the supernatant and keep the pellet. Pat slightly dry on paper
8. dissolve pellet in:
   3 ml Nuclear Lysis Buffer
   50 µl Proteinase K (10 mg/ml)
   300 µl 10% SDS
9. Shake very well and incubate overnight in a water bath at 56°C
10. Add 1 ml 6M NaCl to each tube and shake continuously for 1 min
11. Centrifuge for 20 min at 2 500 x g at room temperature
12. Transfer supernatant to a Falcon tube. Be careful not to transfer any of the pellet or foam. The supernatant must be clear
13. Add 3 volumes ice cold (-20°C) 99.9% ethanol to the supernatant in the Falcon tube and mix very carefully
14. A DNA bundle should form. Carefully hook the bundle out with a needle and place it in an Eppendorf tube that contains 1 ml 70% ethanol
15. Centrifuge at 1 400 x g for 5 min at 4°C
16. Carefully discard the ethanol and allow the pellet to dry
17. Dissolve the pellet in 100-200 µl TE, depending on the size of the pellet

3.2 SureClean Quick-Clean Protocol (Bioline)

1. Add 1 x volume of Quick-Clean to nucleic acid solution, vortex thoroughly and incubate at room temperature for 10 min
2. Centrifuge at maximum speed in a bench-top microfuge for 10 min and discard supernatant
3. Add 100 µl of 70% Ethanol and vortex for 30 sec
4. Centrifuge at maximum speed in a bench-top microfuge for 10 min, remove supernatant and air-dry to ensure complete removal of ethanol
5. Resuspend pellet in desired volume of water

3.3 Big Dye v3.1 Sequencing Chemistry (Applied Biosystems™)

1. Add 4 ng/µl (200-500 bp fragments) or 7 ng/µl (500-1000 bp fragments) of diluted purified PCR product
2. Add 1.3 µl of Big Dye reaction mix
3. Add 2.7 µl Half Dye mix
4. Add 3.3 ρmol/µl of diluted primer

3.4 MSB® Spin PCRapace Columns (Invitek Inc. GmbH)

1. Add 250 µl Binding Buffer to the PCR sample and mix well by vortexing
2. Place a Spin Filter into a 2.0 ml Receiver Tube
3. Transfer the sample completely onto the Spin Filter and centrifuge for 1 min at 12 000 x g
4. Remove filtrate and centrifuge again for 2 min
5. Place the Spin Filter into a new 1.5 ml Receiver Tube
6. Add 10 µl dH_2O directly onto the centre of the Spin Filter
7. Incubate for 1 min at room temperature
8. Centrifuge for 1 min at 10 000 x g

3.5 QIAquick Gel Extraction Kit (Qiagen)

1. Excise the DNA fragment from the agarose gel with a clean, sharp scalpel
2. Weight the gel slice in a colourless tube
3. Add 3 volumes of Buffer QG to 1 volume of gel
4. Incubate at 50°C for 10 min (to help dissolve the gel, mix by vortexing every 2-3 min during the incubation). Solubilize gel completely
5. Add 1 gel volume is isopropanol to the sample and mix
6. Place a QIAquick spin column in a provided 2 ml collection tube
7. To bind the DNA, apply the sample to the QIAquick column, and centrifuge at 13 000 x g for 1 min
8. Discard flow-through and place QIAquick column back in the same collection tube
9. To wash, add 0.75 ml of Buffer PE to QIAquick column and centrifuge at 13 000 x g for 1 min
10. Discard the flow-through and centrifuge the QIAquick column for an additional 1 min at 13 000 x g
11. Place QIAquick column into a clean 1.5 ml microcentrifuge tube
12. To elute DNA, add 30 µl water to the centre of the QIAquick membrane and centrifuge the column at 13 000 x g for 1 min

3.6 *E.cloni*® Chemically Competent Cells (Lucigen Corporation)

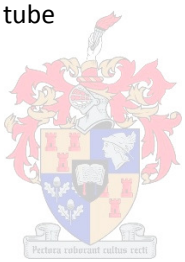
1. For best results, the ligation reaction must be purified or heat killed at 70°C for 15 min before transformation
2. Prepare the YT Agar from the powder included with the cells
3. Chill sterile culture tubes on ice
4. Remove *E.cloni* cells from the -80°C freezer and thaw completely on wet ice (10-20 min)
5. Add 40 µl of *E.cloni* cells to the chilled culture tube
6. Add 1-4 µl of ligation reaction to the 40 µl of cells on ice. Stir briefly with pipette tip
7. Incubate on ice for 30 min
8. Heat shock cells by placing them in a 42°C water bath for 45 sec
9. Return the cells to ice for 2 min
10. Add 960 µl of room temperature Recovery Medium to the cells in the culture tube.
11. Place the tubes in a shaking incubator at 250 rpm for 1 hr at 37°C
12. Plate up to 100 µl of transformed cells on YT agar plates containing the appropriate antibiotic.
13. Incubate the plates overnight at 37°C

3.7 YT Agar Plates

Add the YT Agar powder to 500 ml of deionized water.
YT is per litre: 5 g yeast extract, 8 g tryptone, 5 g NaCl, 15 g agar, pH 7.0

3.8 Genlute Plasmid Mini-prep Kit (Sigma-Aldrich (Pty) Ltd)

1. Pellet cells from 5 ml overnight culture at 12 000 x g for 1 min
2. Discard supernatant
3. Resuspend cells in 200 µl Resuspension Solution
4. Vortex
5. Add 200 µl of Lysis Solution
6. Invert gently to mix
7. Allow to clear for < 5 min
8. Add 350 µl of Neutralization Solution
9. Invert 4-6 times
10. Pellet debris 10 min at max speed
11. Add 500 µl Column Preparation Solution to binding column in a collection tube
12. Spin at 12 000 x g, 1 min
13. Discard flow through
14. Transfer cleared lysate to binding column
15. Spin at 12 000 x g, 1 min
16. Discard flow through
17. Add 750 µl Wash Solution to column
18. Spin at 12 000 x g, 1 min
19. Discard flow through
20. Spin at 12 000 x g, 1 min
21. Transfer column to new collection tube
22. Add 50 µl dH_2O
23. Spin at 12 000 x g, 1 min



APPENDIX 4: REAGENTS AND SOLUTIONS

Reagent	Supplier
Trizma® Base	Sigma-Aldrich (Pty) Ltd.
Boric acid	Sigma-Aldrich (Pty) Ltd.
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich (Pty) Ltd.
Acrylamide	Sigma-Aldrich (Pty) Ltd.
Cresol	Sigma-Aldrich (Pty) Ltd.
Agarose	SeaKem®
Bisacrylamide	Bio Basic Inc.
Amoniumpersulphate	Merck Chemicals (Pty) Ltd.
Sucrose	Merck Chemicals (Pty) Ltd.
Formamide	Merck Chemicals (Pty) Ltd.
Bromophenol blue	Merck Chemicals (Pty) Ltd.
NH_4Cl	Merck Chemicals (Pty) Ltd.
$KHCO_3$	Merck Chemicals (Pty) Ltd.
NaCl	Merck Chemicals (Pty) Ltd.
KH_2PO_4	Merck Chemicals (Pty) Ltd.
TEMED	Fluka Chemie GmbH
Xylene cyanol	Fluka Chemie GmbH
KCl	BDH chemicals Ltd.
Na_2HPO_3	Saarchem (Pty) Ltd.
SDS	BDH Laboratory Supplies
Proteinase K	Finnzymes

All solutions were brought to volume with dH_2O

4.1 Miller *et al.* 1988 DNA Extractions

4.1.1 Lysis Buffer

0.1552 M NH_4Cl
0.0110 M $KHCO_3$
0.0001 M EDTA (pH 8)

4.1.2 Phosphate Buffered Saline (PBS) (pH 7.4)

0.0268 M KCl
0.1369 M NaCl
0.0080 M Na_2HPO_4
0.0015 M KH_2PO_4

4.1.3 Nuclear Lysis Buffer

0.0100 M Tris
0.4004 M NaCl
0.0021 M EDTA (pH 8)

4.1.4 10% Sodium Dodecyl Sulphate (SDS)

0.3468 M SDS

4.2 10X TBE Electrophoresis Buffer (pH 8.3)

0.0890 M Trizma Base
0.0890 M Boric acid
0.0020 M EDTA

4.3 40% Polyacrylamide (PAA), 5% Cross-linkage

5.3461 M acrylamide
0.1297 M bisacrylamide

4.4 15% PAGE Gels

37.5% (w/v) 40% PAA stock
20% (v/v) 5X TBE (pH 8.3)
1% (v/v) 0.1% APS
0.2% (v/v) TEMED

4.5 Cresol Loading Dye

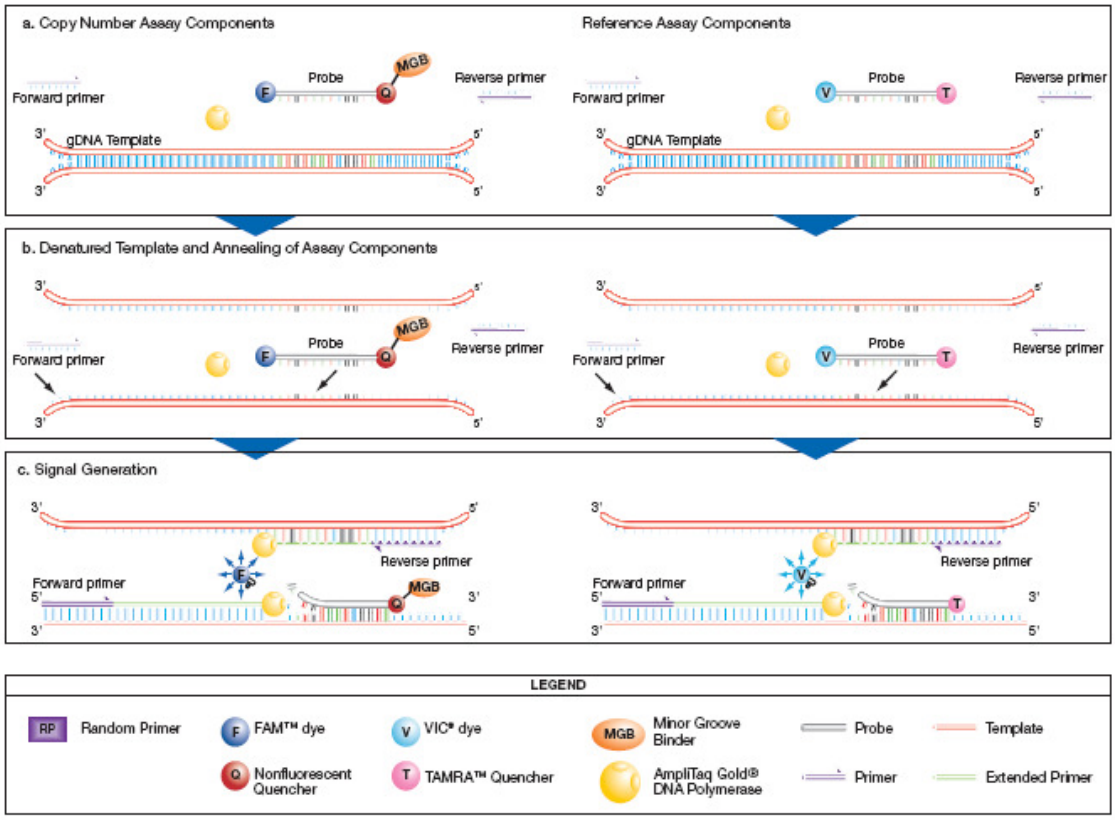
2% (v/v) 10mg/ml cresol stock solution
0.9933 M sucrose

4.6 Bromophenol Blue Loading Dye

90.8% (v/v) formamide
0.0230 M EDTA (pH 8)
0.0009 M xylene cyanol
0.0007 M bromophenol blue

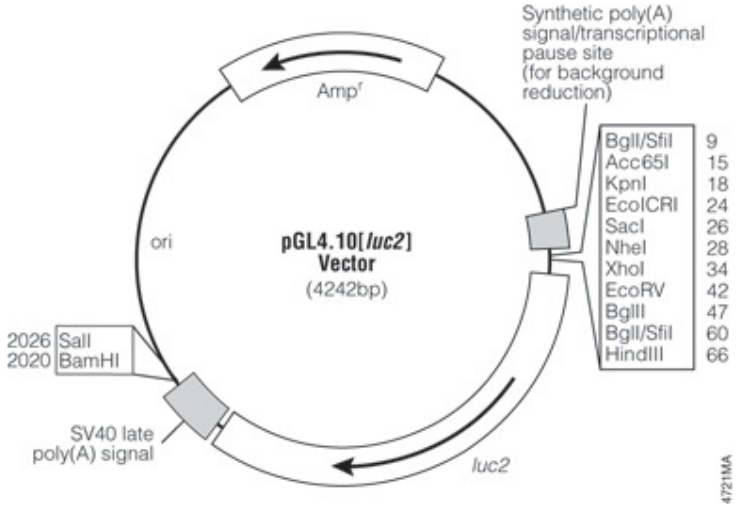
APPENDIX 5: TAQMAN® CNV ASSAY (APPLIED BIOSYSTEMS™)

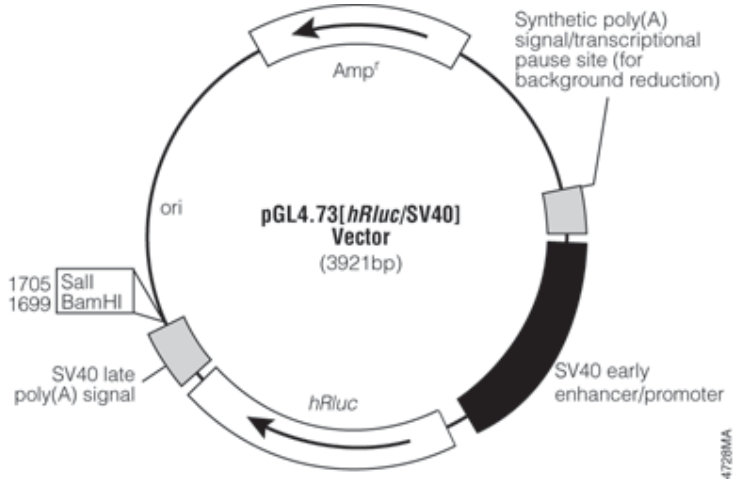
(https://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_062368.pdf)



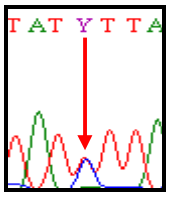
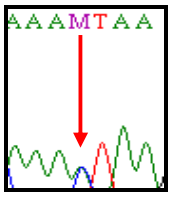
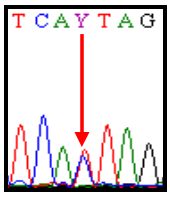
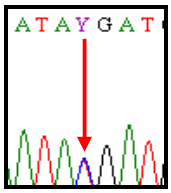
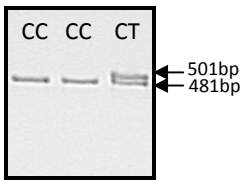
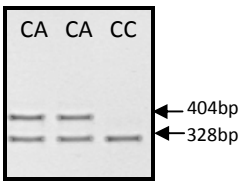
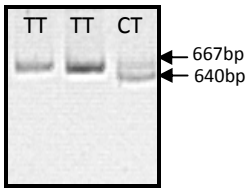
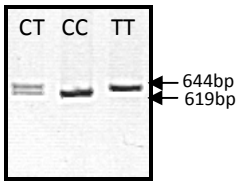
APPENDIX 6: VECTOR MAPS (PROMEGA)

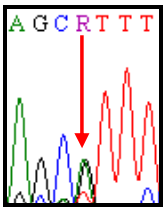
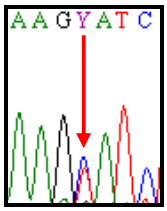
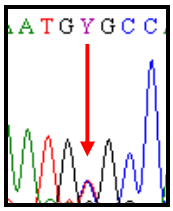
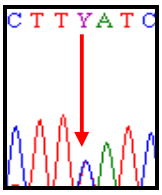
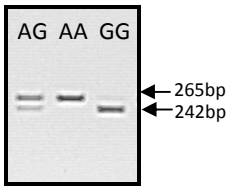
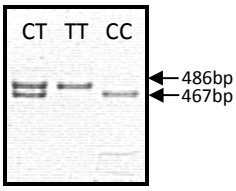
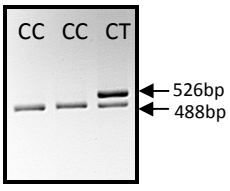
(http://www.promega.com/vectors/allvectors.htm)

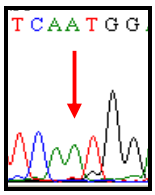
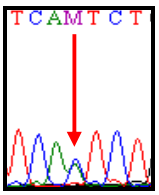
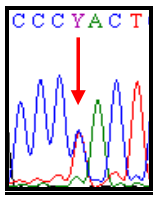
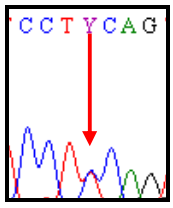
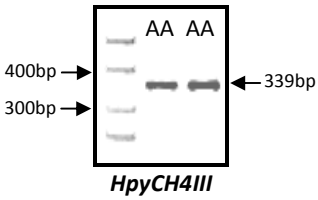
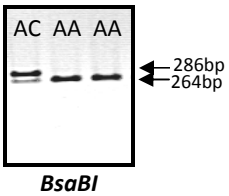
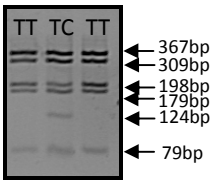


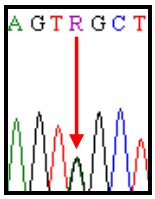
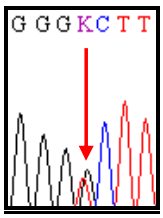
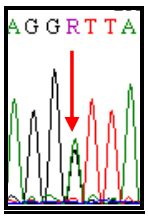
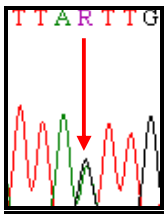
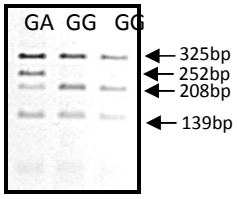


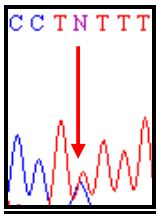
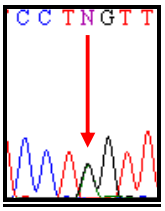
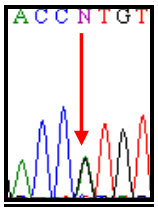
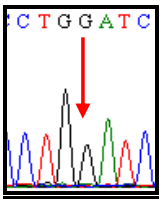
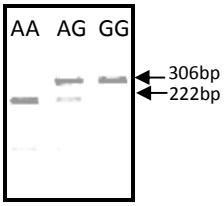
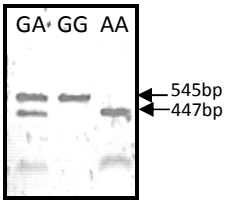
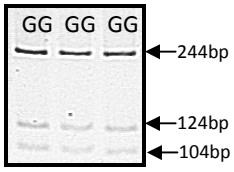
APPENDIX 7: DETECTED VARIANTS

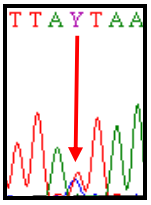
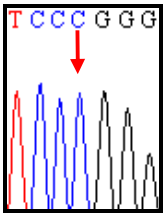
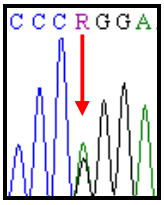
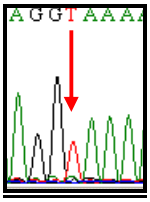
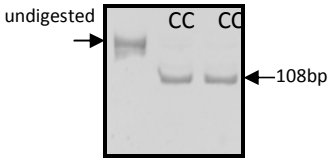
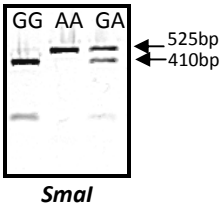
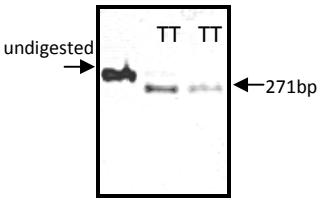
-2030C>T	-2020C>A	-1439T>C rs17878739	-1418C>T rs3814637
			
			
MboI	***DdeI***	***RsaI***	***RsaI***

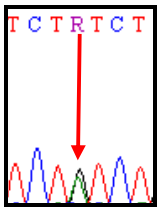
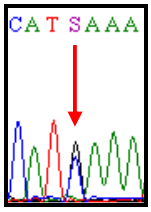
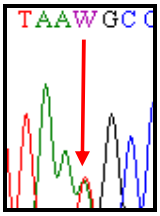
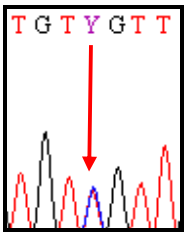
-1041A>G rs7902257	-806C>T rs12248560 (*17)	-783C>T rs11568729	-98T>C rs4986894 (*2C)
			
			Part of *2C*
AclI	***HpyCH4V***	***HhaI***	

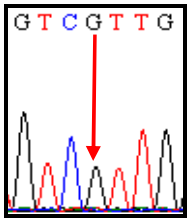
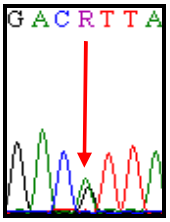
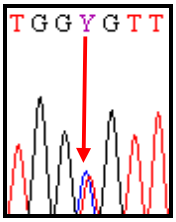
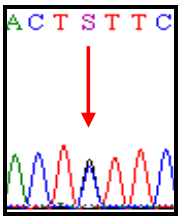
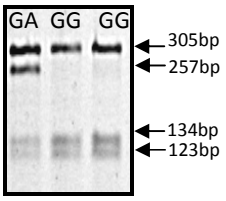
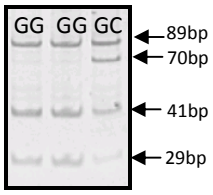
1A>G[1] rs28399504 (*4)	55A>C (I19L) rs17882687 (*15)	99C>T rs17885098 (*1B)	IVS1+11T>C rs17882201
			
		Part of *1B*	
HpyCH4III	***BsaBI***		***BseNI***

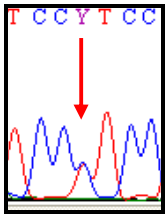
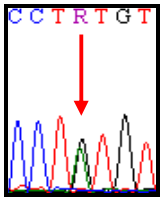
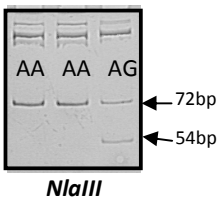
IVS1+20G>A rs17881883	IVS1-340T>G rs17884832 (*3B)	IVS1-231G>A rs7916649 (*2C, *3B)	IVS1-227G>A
			
Intronic	Intronic	Part of *2C*	 ***MseI***

IVS2-48C>T	IVS2-23A>G rs12769205	431G>A (R144H) rs17884712 (*9)	17948G>A[1] rs4986893 (*3)
			
Intronic	 ***BfaI***	 ***HpyCH4III***	 ***BsaJI***

IVS4-40T>C rs57752480	680C>T[1] rs6413438 (*10)	681G>A rs4244285 (*2C)	IVS5+2T>A[1] rs12571421 (*7)
			
Intronic	 ***BslI***	 ***SmaI***	 ***BsaAI***

IVS5+228G>A rs12571421 (*2C)	IVS5-51C>G rs4417205 (*2C)	IVS6-196T>A rs28399513 (*2C)	990C>T rs3758580 (*2C)
			
Part of *2C	Part of *2C	Part of *2C	Part of *2C

991A>G[1] rs3758581 (*1B)	1120G>A (novel V374I)	IVS7-106T>C rs4917623 (*18)	IVS8+122G>C rs17880188
			
Part of *1B	 ***HpyCH4IV***	Intronic	 ***HpyCH4III***

IVS8-119C>T rs12268020	IVS8-17A>G rs4451645
	
Intronic	 ***NlaIII***

[1]These variants were genotyped in the entire cohort, however were not detected in the Xhosa population.

APPENDIX 8: CONFERENCE PRESENTATIONS

Poster Presentations

Drögemöller BI, Malan S, Koen L, Niehaus DJH, Hillermann-Rebello R, Warnich L. Analysis of sequence diversity in the *CYP2C19* gene that could affect oxidative metabolism of therapeutic agents. Faculty of Health Sciences, Stellenbosch University, Neuroscience Annual Academic Day. 13 August 2008. Tygerberg, South Africa.

Drögemöller BI, Malan S, Koen L, Niehaus DJH, Hillermann-Rebello R, Warnich L. Analysis of sequence diversity in the *CYP2C19* gene that could affect oxidative metabolism of therapeutic agents. The 15th Biannual National Congress of the South African Society of Psychiatrists, 10-14 August 2008, Fancourt, George, South Africa.

Drögemöller BI, Malan S, Koen L, Niehaus DJH , Hillermann-Rebello R, Warnich L. Analysis of sequence diversity in the *CYP2C19* gene in a unique South African population. South African Society of Human Genetics Congress. 5-8 April 2009. Spier, Stellenbosch, South Africa.

Wright GEB, Niehaus DJH, **Drögemöller BI**, Koen L, Gaedigk A, Warnich L. Employing a novel genotyping strategy to analyse the *CYP2D6* gene locus in a South African Xhosa schizophrenia population. Pharmacogenomics and Personalized Medicine Meeting. 12-15 September 2009. Hinxton, UK.

Oral Presentations

Drögemöller BI, Malan S, Koen L, Niehaus DJH, Wright GEB, Venter M, Warnich L. From the bench to the bedside: The elucidation of *CYP2C19* sequence diversity for implementation in optimal treatment plans in South African individuals. Biological Psychiatry Congress. 28-31 May 2009. Arabella, Kleinmond, South Africa.

Wright GEB, Niehaus DJH, **Drögemöller BI**, Koen L, Warnich L. Molecular genetic analysis of the *CYP2D6* gene locus in a unique South African population: implications for the pharmacogenetic treatment of Xhosa schizophrenia patients. Biological Psychiatry Congress. 28-31 May 2009. Arabella, Kleinmond, South Africa.

Drögemöller BI, Niehaus DJH, Malan S, Koen L, Wright GEB, Venter M, Warnich L. *CYP2C19* sequence diversity: A missing ingredient in the optimal treatment plans of South African individuals? Faculty of Health Sciences, Stellenbosch University, Neuroscience Annual Academic Day. 12 August 2009. Tygerberg, South Africa.
(Awarded: Best presentation by a young researcher (category <35 years))

Drögemöller BI, Koen L, Niehaus DJH, Malan S, Wright GEB, Venter M, Warnich L. The elucidation of *CYP2C19* sequence diversity in a South African Xhosa population. Pharmacogenomics and Personalized Medicine Meeting. 12-15 September 2009. Hinxton, UK.

APPENDIX 9: MANUSCRIPT TO BE SUBMITTED TO *PHARMACOGENOMICS* (www.futuremedicine.com/loi/pgs)

Characterization of the *CYP2C19* Genetic Profile in two South African Populations

Drögemöller Britt[1], Wright Galen[1], Niehaus Dana[2], Koen Liezl[2], Malan Stefanie[1], Da Silva Danielle[1], Hillermann-Rebello Renate[1], Venter Mauritz[1],Warnich Louise[1]

[1] Department of Genetics, Stellenbosch University, South Africa
[2] Department of Psychiatry, Stikland Hospital, Stellenbosch University, South Africa

Corresponding author: Louise Warnich

Department of Genetics
Stellenbosch University
Private Bag XI
Matieland
7602
South Africa

Telephone number: +27-21-8085888

Fax number: +27-21-8085833

Email address: lw@sun.ac.za

Abstract

Aims: This study was aimed at elucidating common sequence variation present in the *CYP2C19* gene within the South African Xhosa population for future pharmacogenetic applications within the diverse South African populations.

Materials & Methods: The identification of common sequence variation was achieved through sequence analysis of 15 Xhosa individuals. The detected variants were then prioritised for genotyping in an additional 85 Xhosa and 75 Cape Mixed Ancestry (CMA) individuals, while the 5'-upstream variants were prioritised for dual luciferase reporter assays.

Results: Thirty variants were identified in the Xhosa population, including two new *CYP2C19* alleles, *CYP2C19*27* and *CYP2C19*28*, present in both the CMA and Xhosa populations. *CYP2C19*27* is characterized by -1041G>A, which caused a two-fold decrease in luciferase activity, while *CYP2C19*28* is characterized by the non-synonymous, V374I, variant. Among the previously characterised variants, *CYP2C19*2*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17* were present in both populations, while *CYP2C19*3* was only present in the CMA population.

Conclusions: Our data demonstrate that both the Xhosa and CMA populations exhibit unique genetic profiles that could influence the outcome of drug therapy in these South African populations.

Keywords

CYP2C19, Xhosa, Cape Mixed Ancestry, pharmacogenetics, South Africa

Introduction

The frequent occurrence of adverse drug reactions (ADRs) and treatment failure contribute significantly to economic and health-care burdens worldwide, with fatal ADRs falling between the fourth and seventh leading cause of death [1,2] and approximately 50% of patients experiencing ineffective treatment outcomes with most drugs [3]. It has, however, been argued that through the implementation of pharmacogenetics, the rate of ADRs could be reduced by 10-20% and the efficiency of drugs could be increased by 10-15% [4]. To bring to light the South African context, a study by Mehta *et al.* [5], conducted in a hospital in the Cape Town metropolitan area, revealed that 14% of all patients admitted to that hospital exhibited ADRs, with a fatality rate 5-10 fold higher than that reported in USA and UK populations. Furthermore, over half of the ADRs led to hospitalisation and almost a third of hospital-acquired ADRs were preventable, thus emphasising a need to reassess treatment plans, taking pharmacogenetics into consideration.

Bearing the current prevalence of ADRs and treatment failure in mind, the Cytochrome P450 (CYP) genes, which encode phase I drug metabolising enzymes, have been demonstrated to be good candidates for pharmacogenetic studies [6] and may provide information to aid in the attainment of

optimal treatment plans. According to Ingelman-Sundberg *et al.* [7], the *CYP2D6*, *CYP2C9* and *CYP2C19* genes appear to receive the most attention on the Human *CYP* Allele Nomenclature Website [101]. Additionally, members of the *CYP2C* sub-family have been categorised into the pharmacogenetically valid category and are collectively involved in the metabolism of 20% of prescribed drugs [8]. More specifically, the *CYP2C19* gene has been shown to be of particular value in the context of drugs such as antidepressants, anticonvulsants, sedatives, antiulcer, antimalarial, anti-HIV and anti-platelet agents. All of these therapeutic agents are metabolised by the CYP2C19 enzyme and have been associated with ADR and/or treatment failure [9-16]. Thus, the examination of the *CYP2C19* gene is of value to the field of pharmacogenetics.

Several polymorphisms affecting the functioning of the CYP2C19 enzyme have been identified and individuals can be categorised into poor metabolizers (PM), intermediate metabolizers (IM), extensive metabolizers (EM) and ultra-rapid metabolizers (UM) according to the level of enzyme functionality coded by the *CYP2C19* gene. Several populations have been genotyped for the non-functional *CYP2C19*2* and *CYP2C19*3* alleles and differences in the frequencies of these variants have been widely documented [17]. Unfortunately, thorough characterisation of populations to identify novel variants including copy number variants (CNV) as well as other previously identified variants in moderate to large sample sizes remains limited, especially in the genetically diverse African populations. The need for more thorough genotyping panels is well supported by a recent study by Ragia *et al.* [18] performed in a Greek population, where formerly designated *CYP2C19*1* functional alleles were re-genotyped for the *CYP2C19*17* increased function allele. The results from this study showed that the *CYP2C19*1* allele frequency was in fact 44% as opposed to the previously calculated 76%. This further emphasises the need to examine other areas of interest within and around the gene, with special attention to the 5'-upstream genomic areas and their role in controlling the expression of genes.

Differences in allele frequencies between populations have been widely observed [17] and may influence how populations react differently to treatment. It is thus important to examine different populations to determine their unique genetic profiles and assess how these profiles may affect drug metabolism. African populations are of special interest as they are the most ancient of populations and consequently their genomes have been exposed to greater diversification pressures than non-African populations, as reiterated in a recent paper by Tishkoff *et al.* [19]. The different African languages spoken throughout the continent encompass approximately one third of the world's languages, with 2 000 distinct ethno-linguistic groups present [102]. More specifically, the 11 official languages spoken in South Africa, hint at the number of diverse populations residing within the country. Although there remains a vast amount of information to be obtained from these

populations, to our knowledge only one South African population, namely the Venda population, has been examined for *CYP2C19* to date. Two studies have been performed on this population, however the first study examined only the *CYP2C19*2* and *CYP2C19*3* variants [20], while the sample size of the second study was very small and only 313 bp from the transcriptional start codon of the 5'-upstream region were re-sequenced [21]. Additionally, according to the 2001 South African census [103], the Venda comprise only 2.3% of the South African population, making them the second smallest population in the country.

This study places its focus on the Xhosa population, while also examining the Cape Mixed Ancestry (CMA) population. The Xhosa population comprises 17.6% of the South African population, making it the second largest unique South African population, while the mixed ancestry comprises 8.9% of the country's population [103]. The Xhosa, like the Venda, are a Bantu-speaking population belonging to the Niger-Kordofanian African macrofamily [22]. Both the Xhosa and Venda populations exhibit high levels of Southern African Khoisan and Western African Bantu associated ancestral clusters (AAC), as well as low levels of East African Bantu AACs [19] suggesting a similar genetic composition. However, in a more focussed study by Lane *et al.* [23], comparing South African populations, the Venda and Xhosa were shown to cluster apart from one another. Additionally the CMA population is of special interest to population based studies as this population has been shown to exhibit the highest level of intercontinental admixture observed worldwide [19].

Therefore, this study aimed to perform a detailed analysis of the entire *CYP2C19* gene in the Xhosa population, creating a comprehensive *CYP2C19* genetic profile for the Xhosa individuals, which was subsequently used as a guideline for the genotyping of the CMA population. Together these data aim to aid the development of successful pharmacogenetic profiles for other South African populations.

Materials and Methods

Patient Samples

Institutional approval was granted by Stellenbosch University Research Ethics Committee (REC) in terms of the Health Act No 61.2003. For DNA collection purposes; written, informed consent was obtained from 100 Xhosa and 75 CMA control individuals from the Western Cape region. Genomic DNA (gDNA) was extracted from the venous blood using the Miller *et al.* [24] protocol.

Polymerase Chain Reaction (PCR) Amplification and Sequencing Reactions

Primers were designed according to the *CYP2C19* DNA reference sequence (Ensembl Gene ID ENSG00000165841) [104] (Refer to SUPPLEMENTARY TABLE A for primer sequences). PCR amplification

reactions were prepared to a final volume of 25 µl, containing a final concentration of 1X buffer and 0.4 mM dNTPs. All reaction mixes used 15 ng of gDNA, with the exception of the nested PCR reactions which used a 1/100 dilution of PCR product. Excluding the reaction mix for exon 2 and 3, which utilised 0.2 µM of each primer and 1 U of polymerase enzyme, the reaction mixes all contained 0.4 µM of each primer and 0.5 U of polymerase enzyme. All amplification cycle reactions were performed at an initial denaturation of 94°C for 3 min; followed by varying cycles of denaturation at 94°C for 15 sec, annealing for 15 sec and extension at 72°C. All cycles were concluded with a final extension step at 72°C for 5 min (refer to Table 1 for PCR specifications). Subsequent bi-directional semi-automated sequencing was performed using SureClean (Bioline, London, UK), Big Dye v3.1 sequencing chemistry (Applied Biosystems™, Foster City, USA), Half Dye mix (Bioline) and the addition of 0.2% SDS, at a purification cycle of 98°C for 5 min and 25°C for 10 min. Capillary electrophoresis was performed by the Central Analytical Facility of Stellenbosch University on a 3130Xl Genetic Analyzer (Applied Biosystems™). Lastly, the obtained sequencing data were compared to the reference sequence, ENSG00000165841, for the identification of variants.

Table 1: PCR amplification specifications

PCR reagents were supplied by either Bioline, London, UK or Kapa Biosystems, Cape Town, SA.

Region	Products used	$MgCl_2$ conc (mM)	Number of cycles	Annealing temp (°C)	Extension time (sec)
P+E1	Bioline	3.5	40	68	180
E2+3	Bioline	3	40	58	60
E4	Bioline	2	10;30	60;55	30
E5	Bioline	2.5	10;30	65;60	30
E6	Bioline	0.5	40	48	30
E7	Kapa	-	10;30	55;50	30
E8	Kapa	-	10;30	55;50	30
E9	Kapa	-	10;30	60;55	30
3'UTR	Bioline	2	40	62	60
Nested	Bioline	2	25	55	30

P: Promoter, E: Exon, Nested: Nested PCR

Prioritisation and Genotyping of Variants

The identified variants were prioritised for genotyping in an additional 85 healthy Xhosa and 75 CMA individuals according to the following criteria: (i) non-synonymous mutations (using sorting intolerant from tolerant (SIFT) [25], PolyPhen [26] and SNPs3D [105]); (ii) data obtained from the Human *CYP* Allele Nomenclature website [101]; (iii) *in silico* splice site (using NetGene2 [27,28] and SplicePort [29]), transcription factor binding site (using PATCH, [106] MATCH [30] and MatInspector [31]) and mFold analyses [32]; (iv) haplotype tagger data using Haploview v3.31 [33]. In regions which have been deleted without significant effect on expression [34], only those variants creating additional transcription factor binding sites were considered. Those variants not described on the CYP allele

website occurring in perfect linkage disequilibrium (LD) with previously described variants were genotyped in an additional 35 healthy Xhosa individuals. Furthermore, the non-functional *CYP2C19**3 variant was genotyped in the entire Xhosa and CMA cohorts and an additional genotyping protocol was devised to confirm that all rs4244285 (*CYP2C19*2*) variants identified in the Xhosa cohort were not mistyped as a result of the adjacent *CYP2C19*10* (rs6413438) variant. The prioritised variants were subsequently genotyped using restriction fragment length polymorphism (RFLP) analysis, through the preparation of a 20 µl restriction enzyme digestion reaction mix according to the manufacturer's recommendations (refer to SUPPLEMENTARY TABLE B FOR restriction enzyme digest specifications).

TaqMan® Copy Number Assays (Applied Biosystems™)

To detect *CYP2C19* CNVs in the Xhosa cohort, two TaqMan® Copy Number Assay were performed, namely Hs02932336_cn (located in intron 6 of *CYP2C19*) and the TaqMan® Reference Copy Number Assay (*RNase PH1*), in a duplex real-time reaction. The assays were prepared to a final volume of 10 µl according to the manufacturer's instructions in a 384 well plate, using an EpMotion pipetting robot (Eppendorf, Hamburg, Germany), with four replicates for each sample. Thereafter the plate was sealed with optical adhesive film and loaded onto a 7900HT Fast Real-Time PCR system (Applied Biosystems™) at the specified reaction cycle. The results were then analysed on the CopyCaller™ Software (Applied Biosystems™), with the 'Most Frequent Sample Copy Number' set as two and the manual cycle threshold (C_T) set to 0.2.

Plasmid preparation, cell transfection and dual luciferase reporter assays

Dual luciferase reporter assays were used to determine the effect of selected 5'-upstream variants. After initial amplification of the P+E1 fragment, the desired 2 095 bp fragments for insertion into the pGL4.10 vectors (Promega, Madison, USA) were amplified using a nested PCR reaction (Refer to SUPPLEMENTARY TABLE A for primer sequences) with the use of Invitrogen reagents. MSB® Spin PCRapace columns (Invitek Inc. GmbH, Berlin, Germany), QIAquick gel extraction kits (Qiagen, Hilden, Germany), T4 DNA ligase (Invitrogen™, Carlsbad, California, USA), *E.cloni®* Chemically Competent cells (Lucigen Corporation, Middleton, USA), GenElute™ Plasmid Miniprep Kit (Sigma-Aldrich (Pty) Ltd, Aston Manor, South Africa) and HepG2 cells were used to obtain products to measure the luminescence of the constructs on a GloMax™ 96 Plate Luminometer (Promega, Madison, USA). All transfection experiments were performed independently in triplicate to validate the data obtained.

Statistical Analysis

The genotyped SNPs were tested for Hardy-Weinberg equilibrium (HWE) using an analogue to Fisher's exact test in Tools For Population Genetic Analysis (TFPGA) Software v1.3 [35]. The data

obtained from the luciferase assays were tested for normality using a Shapiro-Wilks test and subsequent differences in luciferase activity were tested using a *t*-test. *P* values of <0.05 were considered significant.

Results

Identification of Variants

The re-sequencing data obtained from the 15 Xhosa individuals, revealed 30 variants, all of which were in HWE. Among the detected variants were five novel variants, namely, -2030C>T, -2020C>A, IVS1-227G>A, IVS2-48C>T and 1120G>A (V374I). The novel -2030C>T and -2020C>A variants were found to occur in perfect LD with the *CYP2C19*15* allele-defining variant and the -1439T>C (rs17878739) in the Xhosa and CMA populations. Furthermore, the V374I variant, which resulted in an amino acid change in exon seven, was found to occur in perfect LD with all four of these variants and has been preliminarily designated *CYP2C19*28* [101]. *CYP2C19*28* was detected in both the Xhosa and CMA populations at a frequency of 0.01. In addition to the *CYP2C19*28* allele, the variant -1041G>A was designated *CYP2C19**27 [101] and was found at a frequency of 0.33 in the Xhosa population and 0.08 in the CMA population. The previously characterized *CYP2C19**2, *CYP2C19**9, *CYP2C19*15* and *CYP2C19*17* alleles were detected in the Xhosa population at a frequency of 0.21, 0.09, 0.09 and 0.10, respectively, while they were detected in the CMA population at a frequency of 0.19, 0.04, 0.09 and 0.14. None of the Xhosa individuals tested positive for the *CYP2C19*3*, *CYP2C19*10* or copy number variants, however the *CYP2C19*3* variant was detected in the CMA population at a frequency of 0.07. After classification of the 100 Xhosa individuals according to metabolizer class, it was observed that PMs (individuals with two *CYP2C19*2*/**3* alleles), IMs (individuals with one *CYP2C19*2*/3 allele or one *CYP2C19*9* allele), EMs and UMs (individuals with *CYP2C19*17* alleles unaccompanied by *CYP2C19*2*/*3/**9* alleles) occurred at a frequency of 0.03, 0.49, 0.39 and 0.09, respectively, while the frequencies in the CMA population were 0.08, 0.40, 0.35 and 0.17. A complete table displaying the variants detected and the corresponding frequencies for both the Xhosa and CMA populations can be viewed in SUPPLEMENTARY TABLE C.

In silico analysis

The splice site, mFold and non-synonymous mutation analysis did not reveal any significant results for any of the previously uncharacterized variants. Transcription factor binding site analysis predicted the removal of a GATA-factor binding site as a result of the -2030C>T variant and the addition of an octamer binding protein-1 (Oct-1) binding site, as a result of the -1041G>A variant.

Dual luciferase reporter assays

Statistical analysis showed that there was a significant decrease in the fold induction of the construct containing the *CYP2C19*27* 5'-upstream region (-1041A) when compared to the construct containing

the *CYP2C19*1* 5'-upstream region (-1041G) (*P*=0.0077) (refer to Figure 1). Although the construct containing the *CYP2C19*28* 5'-upstream variants (-2030T variant and those variants occurring in phase with this variant), showed a trend towards decreased luciferase activity when compared to the *CYP2C19*1* construct, no significant difference in fold induction was observed (*P*=0.0928) (refer to Figure 1).

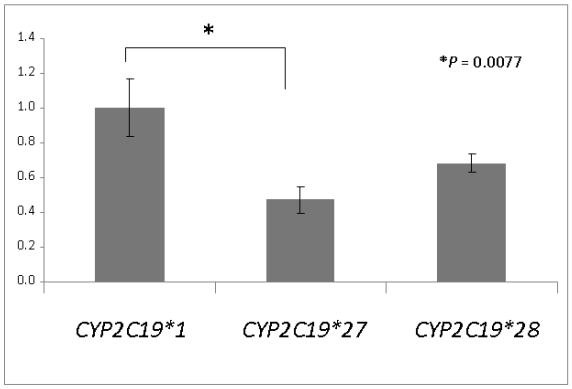


Figure 1: Fold induction ± SEM of constructs containing *CYP2C19*1*, *CYP2C19*27* and *CYP2C19*28* 5'-upstream regions.

Discussion

This study focused on elucidating *CYP2C19* sequence variation in the South African Xhosa population. Novel variants were identified and it was revealed that both the Xhosa and CMA populations exhibit unique genetic profiles, differing from other populations examined to date. Where most studies have examined only the *CYP2C19*2* and *CYP2C19*3* variants, both of which are genotyped by the first FDA approved pharmacogenetics tests, the Roche AmpliChip [107], this analysis has revealed the presence of other relevant variants in both the Xhosa and CMA populations, which could influence the phenotype of patients. These data emphasize the requirement for appropriate genotyping platforms in these and other South African populations.

The novel *CYP2C19*28*, V374I, does not appear to exert a major effect on the CYP2C19 enzyme as none of the algorithms used, predicted a significant impact on the protein structure and to our knowledge there is no known data indicating that the substituted amino acid plays an essential role in the active site of the protein. However, before final conclusions can be made, genotype-phenotype studies are required.

The *CYP2C19*27* allele was shown to result in a decrease in luciferase activity. This finding may be attributed to the predicted addition of an Oct-1 binding site due to the -1041G>A variant, as Oct-1 has been shown to repress the expression of other *CYP* genes [36,37]. Considering the relatively high

frequency of -1041A in the Xhosa population (refer to Figure 2), these data are particularly important to South African populations. Although preliminary bioinformatic analysis suggested the removal of a GATA factor binding site as a result of the novel -2030C>T variant and GATA factors have been implicated in *CYP* expression [38], these results were not validated by the dual luciferase reporter assays. It is important to bear in mind that the other *CYP2C19*28* variants occurring alongside the -2030T variant could have influenced the results obtained for that construct. Additionally, other factors such as additional GATA factors within the 5' region may compensate for the disruption of this GATA transcription factor binding site. While no direct conclusions can be drawn prior to phenotypic validation of these results, this study provides a platform from which to commence these genotype-phenotype correlation studies.

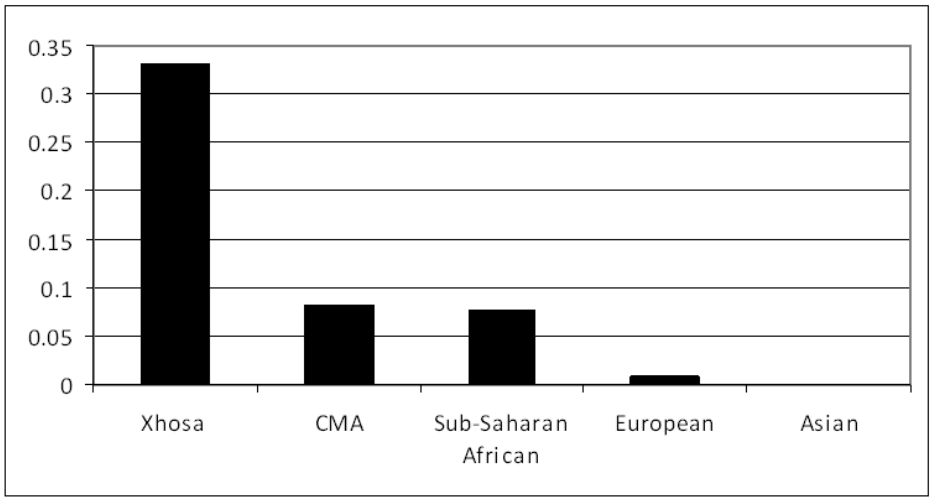


Figure 2: Frequency of the -1041G>A variant in the Xhosa and CMA populations when compared to data obtained from NCBI for the European, Asian and Sub-Saharan populations [108].

Another 5'-upstream variant of importance, detected in this study, was the *CYP2C19*17* (rs12248569) variant. To our knowledge this is the first study to examine and report the presence of this variant in a South African population. Although this variant has been shown to affect the plasma concentration of Escitalopram, with homozygous *CYP2C19*17* individuals exhibiting a 42% lower plasma concentration than homozygous *CYP2C19*1* individuals [39], it has remained largely neglected in most studies. The results of this study suggest that the approach used by Ragia *et al.* [18] should be replicated in other studies to re-evaluate the genotypes of previously examined African populations and thereby create more comprehensive profiles for *CYP2C19* in these populations.

When the frequency of variants present in the Xhosa population were compared to the South African Venda [20,21] population, the degree of similarities observed between the two populations correlates with the population structures reported by Lane *et al.* [23]. It must be noted that accurate

comparisons can only be made using the data obtained from the study by Dandara *et al.* [20], which utilized a similar sample size to this study (75 Venda individuals vs 100 Xhosa individuals) (refer to Figure 3A). The differences and similarities observed between these two South African Bantu populations are of particular interest to South African pharmacogenetic studies, as although they demonstrate that the data obtained from one South African Bantu population may be used to guide the treatment of another Bantu population, these populations may nonetheless need to be treated independently.

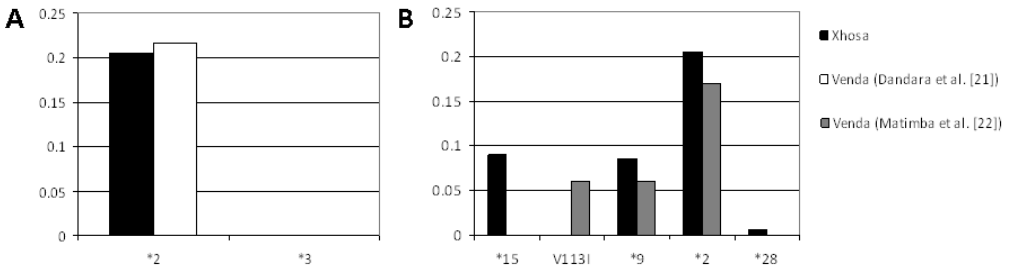


Figure 3: Frequency comparisons between the main variants identified in the Venda and Xhosa populations. Figure 4A demonstrates the similar frequencies observed for *CYP2C19*2* and *CYP2C19*3* in both populations [20], while Figure 4B depicts the discrepancies observed. The differences in allele frequencies depicted in Figure 4B may be attributed to the small sample size (9 Venda individuals vs 100 Xhosa individuals) used in the study performed by Matimba *et al.* [21].

Although *CYP2C19*10* (rs6413438) has been reported to in African-American populations [40] and may thus contribute to the erroneous genotyping of *CYP2C19*2* in other populations of African descent, the re-examination of all individuals genotyped for *CYP2C19*2* in this study did not reveal the presence of the *CYP2C19*10* allele in the Xhosa population. Although *CYP2C19*10* and *CYP2C19*2* appear to exhibit similar enzyme activities and their misclassification does not therefore appear to impact pharmacogenetic applications extensively [41], these data serve as valuable reminders that data from studies involving African-American individuals cannot necessarily be inferred on other African populations.

With regards to the CMA population, all variants genotyped, including the novel and *CYP2C19*3* variants, were present in the population, making this is the first study to our knowledge to detect the *CYP2C19*3* variant in a South African population. The presence of a wide variety of variant points to the high level of admixture from the European, South Asian, Indonesian and Xhosa populations [42] in the CMA population and suggests that additional variants which have previously been identified in other populations, may be detected. In addition to this, novel variants may be identified in the *CYP2C19* gene in this population, as was observed in the *CYP2D6* gene in a study by Gaedigk and Coetsee [43]. The high level of admixture observed in the CMA population complicates the implementation of pharmacogenetic technologies in this population in a manner akin to the complications encountered in other African populations, due to the high level of genetic diversity.

Therefore this population, like other South African populations will, in the future, require more comprehensive analysis in combination with carefully designed genotyping strategies.

When comparing the metabolizer classes of the Xhosa to the Caucasian, Asian and CMA populations, the Asians showed the highest frequency of PMs, the Xhosa showed the highest frequency of IMs and the Caucasians showed the highest frequency of UMs. The CMA population appears to form intermediate frequencies of the metabolizer groups (refer to Figure 4), reflecting the high level of admixture observed in this population. The difference in frequencies of metabolizer classes observed between the Xhosa and CMA populations serves once again as a reminder that the different populations residing in South Africa may need to be considered independently. Furthermore, it should be noted that the metabolizer classes reported in these two examined South African populations do not take any novel/uncharacterised variants into account, therefore after phenotypic validation of these uncharacterised variants, the frequencies of metabolizer classes may change. It must be taken into account that the Caucasian population was not genotyped for *CYP2C19*9* which contributed to the number of IMs observed in both the Xhosa and CMA populations, however, this allele has been reported to occur predominantly in individuals of African descent [40], thus it is not likely to contribute significantly to the number of IMs observed in the Caucasian population.

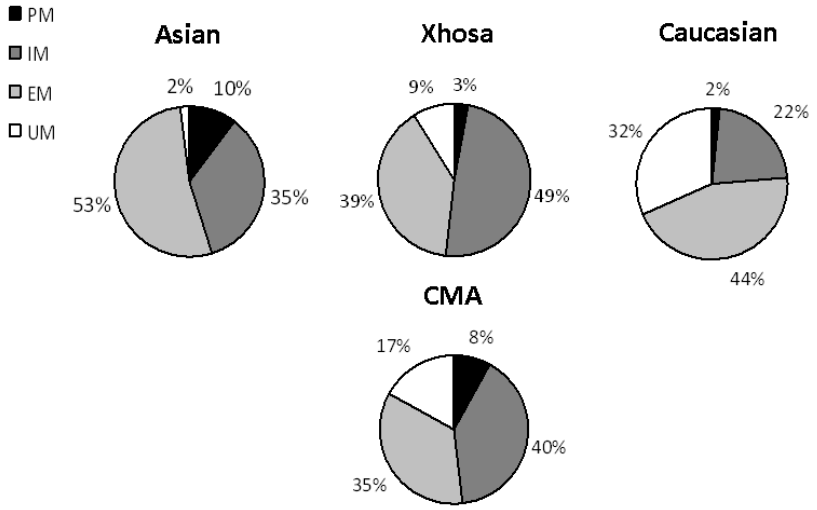


Figure 4: Differences in the frequencies of metabolizer classes observed in the Xhosa, Caucasian, Asian and CMA populations (Ragia *et al.* [18], Chen *et al.* [44]).

Considering that at least 11.6% of the South African population is infected with HIV/AIDS [109], the high frequency of IM individuals is of significance to the country, as studies focusing on *CYP2D6* have noted that IM status may shift towards PM status in HIV/AIDS infected individuals [45], while the CYP activity of HIV/AIDS infected individuals has been shown to be reduced [46]. Furthermore, it has been noted that in patients younger than 60 years, those infected with the HIV virus are twice as

likely to develop ADRs when compared to those unaffected by the disease, while individuals receiving antiretrovirals (ART) are 10 times more likely to develop ADRs than those not taking ARTs [6]. This is importance since CYP2C19 has been shown to be involved with the metabolism of ARTs [47].

Conclusions

The data obtained from this study demonstrate that the *CYP2C19* genetic profile of the Xhosa population differs to that of the Caucasian, Asian, CMA and other African populations. This emphasizes the need to determine the genetic profiles of other South African populations in order to obtain comprehensive pharmacogenetic guidelines for successful implementation in treatment plans within the South African context. This is especially important when considering the high level of variation observed in African populations when compared to non-African populations [19] as emphasized by this study.

It is clear that in the context of HIV/AIDS combined with the high level of ADRs observed in South Africa, revision of treatment plans is required. Systems to optimize treatment plans and thereby decrease unnecessary expenditures on ADRs and redundant treatments are required. To our knowledge this is the most extensive study performed on the *CYP2C19* gene in a South Africa population and luciferase assays have revealed a variant which may affect CYP2C19 enzyme expression. The data obtained in this study have shown that the *CYP2C19* gene in this population contains variation that will most likely affect the metabolism of drugs in this population; therefore after phenotypic validation, these data could play a valuable role in the optimization of treatment plans within South Africa.

Future Perspective

With sequencing technology improving at an exponential rate, allowing for quicker and cheaper analysis, the challenges which are associated with the highly diverse African genomes are gradually lessening, allowing for the identification of rare variants contributing to poor enzyme functioning. The release of the 1000 Genomes Project should aid in the identification of rare variants, with the corresponding pharmacogene database [111] allowing for easy examination of Very Important Pharmacogenes, including *CYP2C19*. Unfortunately, to date Southern African populations are not extensively represented. It has been shown that there are more differences observed between the genomes of two Khoisan individuals than the differences observed between the genomes of Asian and European individuals [48], therefore further sequencing data from these genetically diverse populations will allow for a more complete picture of pharmacogenetic variation. These data should aid in the development of population-based pharmacogenetic approaches in third world countries such as South Africa, where although optimised treatment plans are urgently required, the lack of

resources presently encumbers the implementation of individualised treatment in the crowded and understaffed health care systems.

As genetic variation is not the sole contributor to poor response in patients receiving medication, in future, applicable algorithms such as those described for warfarin [49], are likely to be developed to allow for the calculation of the correct dosage and drug to administer, based on among other things; genotype, physiological/disease status, environmental factors and concomitant drugs. This may be useful in African countries with regards to individuals infected with HIV/AIDS, as it has been reported that the CYP enzyme activity of infected individuals is significantly altered [45]. Therefore considering the high frequency of IMs within the South African Xhosa population, the treatment plans of IM individuals infected with HIV/AIDS may require re-evaluation. Additionally, non-traditional genetic variation such as copy number variants, epigenetics and expression profiling will need to be further considered.

As technology advances and interdisciplinary approaches, education and accessibility are improved, genotype-based treatments will hopefully be implemented into the clinical setting within South Africa. There does, however, remain a lot of work to be done and clinical validation is required before clinicians can realistically consider the implementation of genetic tests as part of their diagnosis and prescriptions.

Executive Summary

Identification of Variants

Thirty variants were identified in the Xhosa population, including five novel variants.

Detection of *CYP2C19** allele

Two new alleles, designated *CYP2C19*27* and *CYP2C19*28*, were detected in both the Xhosa and CMA populations. Additionally, the previously identified *CYP2C19*2*, *CYP2C19*9*, *CYP2C19*15* and *CYP2C19*17* variants were detected in both populations, while *CYP2C19*3* was only detected in the CMA population.

Dual Luciferase Reporter Assays

The -1041G>A variant (*CYP2C19*27*) was shown to cause a two-fold decrease in luciferase activity.

Conclusions

The South African populations examined, displayed unique genetic profiles which are likely to affect drug metabolism. It remains to be determined whether the high level of IMs in these populations contributes to the high level of ADRs observed in individuals infected with HIV/AIDS.

Financial Disclosure/Acknowledgements

We would like to acknowledge the Harry Crossley Foundation and South African National Research Foundation for financial assistance, the Xhosa individuals for their participation, Ms M Bosman for technical assistance, Mr A La Grange for statistical analysis and the Central Analytical Facility of Stellenbosch University.

References

1. Lazarou J, Pomeranz BH, Corey PN: Incidence of adverse drug reactions in hospitilazed patients: a meta-analysis of prospective studies. *JAMA* 279, 1200-1205 (1998).
2. Wester K, Jonnson AK, Sigset O, Druid H, Hagg S: Incidence of fatal adverse drug reactions: a population based study. *Br. J. Clin. Pharmacol.* 65, 573-579 (2008).
3. Allison M: Is personalized medicine finally arriving? *Nat. Biotechnol.* 26, 509-517 (2008).
4. Ingelman-Sundberg M: Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, the present and future. *Trends Pharmacol. Sci.* 25, 193-200 (2004).
5. Mehta U, Durrheim DN, Blockman M, Kredo T, Gounden R, Barnes KI: Adverse drug reactions in adult medical inpatients in a South African hospital serving a community with a high HIV/AIDS prevalence: prospective observational study. *Br. J. Clin. Pharmacol.* 65, 396-406 (2007).

• Highlighting the high prevalence of ADRs in South Africa influenced among other things by the HIV/AIDS epidemic. This article provides evidence that better treatment plans are required in South Africa.

6. Oscarson M: Pharmacogenetics of drug metabolising enzymes: importance for personalised medicine. *Clin. Chem. Lab. Med.* 41, 573-580 (2003).
7. Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C: Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol. Ther.* 116, 496–526 (2007).
8. Goldstein JA: Clinical relevance of genetic polymorphisms in the human CYP2C subfamily. *Br. J. Clin. Pharmacol.* 52, 349–355 (2001).
9. Xie HG, Kim RB, Stein CM, Wilkinson GR, Wood AJJ: Genetic polymorphism of (S)-mephenytoin 4'hydroxylation in populations of African descent. *Br. J. Clin. Pharmacol.* 48, 402–408 (1999).
10. Sagar M, Janczewska I, Ljungdahl A, Bertilsson L, Seensalu R: Effect of *CYP2C19* polymorphism on serum levels of vitamin B12 in patients on long-term omeprazole treatment. *Aliment. Pharmacol. Ther.* 13, 453-458 (1999).
11. Herrlin K, Yasui-Furokori N, Tybring G, Widen J, Gustafsson LL, Bertilsson L: Metabolism of citalopram enantiomers in CYP2C19/CYP2D6 phenotyped panels of healthy Swedes. *Br. J. Clin. Pharmacol.* 56, 415–421 (2003).
12. Kirchheiner J, Brosen K, Dahl ML *et al.*: *CYP2D6* and *CYP2C19* genotype-based dose recommendations for antidepressants: a first step towards subpopulation-specific dosages. *Acta. Psychiatr. Scand.* 104, 173-192 (2001).
13. Inomata S, Nagashima A, Itagaki F *et al.*: *CYP2C19* genotype affects diazepam pharmacokinetics and emergence from general anesthesia. *Clin. Pharmacol. Ther.* 78, 647–655 (2005).
14. Ford GA, Wood SM, Daly AK: *CYP2D6* and *CYP2C19* genotypes of patients with terodiline cardiotoxicity identified through the yellow card system. *Br. J. Clin. Pharmacol.* 50, 77-80 (2000).
15. Gardiner SJ, Begg EJ: Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. *Pharmacol. Rev.* 58, 521–590 (2006).

16. Shuldiner AR, O'Connell JR, Bliden KP *et al.*: Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficiency of clopidogrel therapy. *JAMA* 302, 849-857 (2009).
17. Sistonen J, Fuselli S, Palo JU, Chauhan N, Padh H, Sajantila A: Pharmacogenetic variation at *CYP2C9*, *CYP2C19* and *CYP2D6* at global and microgeographic scales. *Pharmacogenet. Genomics* 19, 170–179 (2009).

• A good review summarizing the varying frequencies of *CYP2C19* alleles in individuals of difference ethnicities.

18. Ragia G, Arvanitidis KI, Tavridou A, Manolopoulos VG: Need for reassessment of reported *CYP2C19* allele frequencies in various populations in view of *CYP2C19*17* discovery: the case of Greece. *Pharmacogenomics* 10, 43-49 (2009).

• Demonstrating the importance of genotyping *CYP2C19*17* in populations and thus the need for more thorough genotyping panels.

19. Tishkoff SA, Reed FA, Friedlaender FR *et al.*: The genetic structure and history of Africans and African Americans. *Science* 324, 1035-1044 (2009).

•• Evidence for the high level of diversity in African genomes and a need for more thorough genetic analysis of these populations.

20. Dandara C, Masimirembwa CM, Magimba A *et al.*: Genetic polymorphism of *CYP2D6* and *CYP2C19* in East- and Southern African populations including psychiatric patients. *Eur. J. Clin. Pharmacol.* 57, 11-17 (2001).
21. Matimba A, Del-Favero J, Van Broeckhoven C, Masimirembwa C: Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects. *Hum. Genomics* 3, 169-190 (2009).

• An indication of which *CYP2C19* genotypes could be expected in African populations as well as proof that these populations differ from one another.

22. Ehret C: Bantu expansions: re-envisioning a central problem of early African history. *Int. J. Afr. Hist. Stud.* 34, 5-41 (2001).
23. Lane AB, Soodyall H, Arndt S *et al.*: Genetic substructure in South African Bantu-speakers: evidence from autosomal DNA and Y-chromosome studies. *Am. J. Phys. Anthropol.* 119, 175-185 (2002).
24. Miller SA, Dykes DD, Polesky HF: A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16, 1215-1215 (1988).
25. Ng PC, Henikoff S: SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812-3814 (2003).
26. Ramensky V, Bork P, Sunyaev S: Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900 (2002).
27. Brunak S, Engelbrecht J, Knudsen S: Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220, 49-65 (1991).
28. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* 24, 3439-3452 (1996).
29. Dogan RI, Getoor L, Wilbur WJ, Mount SM: SplicePort: an interactive splice-site analysis tool. *Nucleic Acids Res.* 35, 285-291 (2007).
30. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31, 3576-3579 (2003).
31. Quandt K, Frech K, Karas H, Wingender E, Werner T: MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* 23, 4878-4884 (1995).
32. Zuker M: Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406-3415 (2003).
33. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-265 (2005).

34. Arefayene M, Skaar TC, Zhao X *et al.*: Sequence diversity and functional characterization of the 5'-regulatory region of human *CYP2C19*. *Pharmacogenetics* 13, 199-206 (2003)
35. Miller M: Tools for population genetic analyses (TFPGA) 1.3: a Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by the author (1997).
36. Bhat R, Weaver JA, Sterling KM, Bresnick E: Nuclear transcription factor Oct-1 binds to the 5' upstream region of CYP1A1 and negatively regulates its expression. *Int. J. Biochem. Cell Biol.* 28, 217-227 (1996).
37. Fiala-Beer E, Lee AC, Murray M: Regulation of the rat CYP2A4 gene promoter by c-Jun and octamer binding protein-1. *Int. J. Biochem. Cell. Biol.* 39, 1235-1247 (2007).
38. Thum T, Haverich A, Borlak J: Cellular dedifferentiation of endothelium is linked to activation and silencing of certain nuclear transcription factors: implications for endothelial dysfunction and vascular biology. *Faseb J.* 14, 740-751 (2000).
39. Rudberg I, Mohebi B, Hermann M, Refsum H, Molden E: Impact of the ultrarapid CYP2C19*17 allele on serum concentration of escitalopram in psychiatric patients. *Clin. Pharmacol. Ther.* 83, 322-327 (2008).
40. Blaisdell J, Mohrenweiser H, Jackson J *et al.*: Identification and functional characterization of new potentially defective alleles of human CYP2C19. *Pharmacogenetics* 12, 703-711 (2002).
41. Rasmussen H, Werge T: Misclassification of allele *CYP2C19*10* as *CYP2C19*2* by a commonly used PCR-RFLP procedure. *Genet. Test.* 12, 57-58 (2008).
42. Patterson N, Petersen DC, van der Ross RE *et al.*: Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* doi:10.1093/hmg/ddp505 (2009).
43. Gaedigk A, Coetsee C: The *CYP2D6* gene locus in South African Coloureds: unique allele distributions, novel alleles and gene arrangements. *Eur. J. Clin. Pharmacol.* 64, 465-475 (2008).
44. Chen L, Qin S, Xie J *et al.*: Genetic polymorphism analysis of CYP2C19 in Chinese Han populations from different geographic areas of mainland China. *Pharmacogenomics* 9, 691-702 (2008)
45. O'Neil WB, Gilfix BM, Markoglou N, Di Girolamo A, Tsoukas CM, Wainer IW: Genotype and phenotype of cytochrome P450 2D6 in human immunodeficiency virus-positive patients and patients with acquired immunodeficiency syndrome. *Eur. J. Clin. Pharmacol.* 56, 231-240 (2000).
46. Jones AE, Brown KC, Werner RE *et al*.: Variability in drug metabolizing enzyme activity in HIV-infected patients. *Eur. J. Clin. Pharmacol.* DOI 10.1007/s00228-009-0777-6 (2010).
47. Haas DW, Smeaton LM, Shafer RW *et al.*: Pharmacogenetics of long-term responses to antiretroviral regimens containing efavirenz and/or nelfinavir: an adult aids clinical trials group study. *J. Infect. Dis.* 192, 1931−1942 (2005).
48. Schuster SC, Miller W, Ratan A *et al* : Complete Khoisan and Bantu genomes from southern Africa. *Nature*. doi:10.1038 (2010).

•• Highlights the vast amount of data to be obtained from Southern African genomes and a need for further studies to be performed.

49. Sasaki T, Tabuchi H, Higuchi S, Ieiri I: Warfarin-dosing algorithm based on a population pharmacokinetic/pharmacodynamic model combined with Bayesian forecasting. *Pharmacogenomics*. 10, 1257-1266 (2009)

Websites:

101. Home page of the human cytochrome P450 (CYP) allele nomenclature committee http://www.cypalleles.ki.se

102. Ethnologue
www.ethnologue.com
103. Statistics South Africa (Census 2001).
http://www.statssa.gov.za/census01/HTML/default.asp

104. Ensembl
http://apr2007.archive.ensembl.org/index.html
105. SNPs3D
http://www.snps3d.org/
106. Pattern Search for Transcription Factor Binding Sites
http://www.gene-regulation.com/cgi-bin/pub/programs/patch/bin/patch.cgi
107. AmpliChip
http://www.amplichip.us/
108. Nucleotide Polymorphism Database (PrimerBlast)
http://www.ncbi.nlm.nih.gov/tools/primer-blast/

109. UNAIDS. South Africa
http://www.unaids.org/en/Regions_Countries/Countries/south_africa.asp

Supplementary Material:

Supplementary Table A: Primer sequences

Region	Template Primers (5' – 3')		Internal Primers (5' – 3')	
P+E1	CYP2C19P+E1:F	CAG AAC TGG AAC ACC TAG CTC TCA		
P+E1	CYP2C19P+E1:R	GAC AGA CTG GAA AAG GCA ACA AAA G		
P			CYP2C19P+E1:R1	CTC ACA CCT CAT ATC CCT TTG GAA TCT CTC
P			CYP2C19P+E1:R2	GAG ATG CTT TGA GAA CAG AAG ACA C
P			CYP2C19P+E1:F3	GTG TCT TCT GTT CTC AAA GCA TCT C
P			CYP2C19P+E1:FS1	GGA CAA AGT CTC CTA ATC TTC GAT ATA G
P			CYP2C19P+E1:RS1	CAC CGT CAT AAT TGA GAG CAC TGA AG
P			CYP2C19P+E1:mF2020	CTA AAG AGA GCA ACC AAG CT**g** AT
P			CYP2C19-1439mF	CTT AAT AAG AGA ACT GGA AAT AAC C**gt** A
P			CYP2C19-1418mF	CCT CAT TAG GAA ATT TAG AAC AA**g** TA
P			CYP2C19-1041mF	GCT CTT CCT TCA GTT ACA CTG A**a**C
P			CYP2C19P+E1:mF*17	GTG TCT TCT GTT CTC Aa**t** G
E1			CYP2C19P+E1:mF*15	CTC TCA TGT TTG CTT CT**g a**TT TCA
E2+3	CYP2C19E2+3:F	CAT AAA AGA CTG TTG GAC CAG G		
E2+3	CYP2C19E2+3:R	AGG AGA GCA GTC CAG AAA GG		
E3			CYP2C19E2+3:FS1	GCA CAC CTA CCA AAT CCT CTG
E2			CYP2C19E2+3:RS1	GGA GAT CCC AGG CAA GAA AGA GG
E4	CYP2C19E4:F	GCA ACC ATT ATT TAA CCA GCT AGG		
E4	CYP2C19E4:R	TCA AAA ATG TAC TTC AGG GCT TGG TC		
E5	CYP2C19E5:F	CAA CCA GAG CTT GGC ATA TTG T		
E5	CYP2C19E5:R	GCA GAA CAG AGC TTT TCC TAT C		
E5			CYP2C19E5:mF*10	GGT TTT TAA GTA ATT TGT T**a**c **c**GG TTC C
E6	CYP2C19E6:F	GCA TTC CCT TTG AAA ACT GGC ACA AGA C		
E6	CYP2C19E6:R	CAC ACC ATT AAA TTG GGA CAG ATT ACA GC		
E7	CYP2C19E7:F	GGG CTT CTC TTC CTT CTT TCA TTT CT		
E7	CYP2C19E7:R	CTC TCA CCC AGT GAT GGT AGA GGG'		
E8	CYP2C19E8:F	CGT CTA TCT GTC TGG AAA TGG'		
E8	CYP2C19E8:R	GAG GAT GTA TCA CCA GCG		
E9	CYP2C19E9:F	CAC CCA TCC ATC CTT TCA TTC ATG C'		
E9	CYP2C19E9:R	GGA CCA GAG GAA AGA GAG CTG	CYP2C19E9:mF	CAC ATG AGG AGT AAC TTC TCC **a**T
E9			CYP2C193UTR:F	CAC ATG AGG AGT AAC TTC TCC CT
3'UTR	CYP2C193UTR:R	CCT CAT GTA ACT CTA AAT TTT GG		

P: Promoter, E: Exon, F: Forward primer, R: Reverse primer, S: Sequencing primer, m: Mutagenic primer, Lower case letters: Mutagenic bases

Region	Dual Luciferase Reporter Assay Primers (5' – 3')	
P	CYP2C19P:F	CCC CC**c tcg ag**C AGA ACT GGA ACA CCT AGC TCT C
P	CYP2C19P:R	CCC CC**a gat ct**C TCAC ACC TCA TAT CCC TTT GG

P: Promoter, F: Forward primer, R: Reverse primer, Lower case bold letters: *XhoI* and *BglII* recognition sites

Supplementary Table B: RFLP specifications (Enzymes were supplied by New England Biolabs Inc., Beverly, USA)

Variant	Allele	Restriction Enzyme	Temperature (ºC) and Additives	Genotype	Size of Fragments (bp)	Gel	Primer set
-2030C>T	Novel	*Mbo*I	37	CC CT TT	481, 20 501, 481, 20 501	2% Agarose	CYP2C19P+E1:mF2020 CYP2C19P+E1:R1
-2020C>A	Novel	*Dde*I	37	CC AC AA	328, 76, 66, 57, 16 404, 328, 76, 66, 57, 16 404, 66, 57, 16	2% Agarose	CYP2C19P+E1:F CYP2C19P+E1:R1
rs17878739		*Rsa*I	37	TT TC CC	667 667, 660, 27 660, 27	3% Agarose	CYP2C19-1439mF CYP2C19P+E1:R2
rs3814637		*Rsa*I	37	CC CT TT	619, 25 644, 619, 25 644	3% Agarose	CYP2C19-1418mF CYP2C19P+E1:R2
rs7902257	**27*	*Acl*I	37	GG AG AA	265 265, 242, 23 242, 23	3% Agarose	CYP2C19-1041mF CYP2C19P+E1:R2
rs17882201		*BseN*I	65	TT CT CC	367, 309, 198, 179, 79, 9 367, 309, 198, 179, 124, 79, 74, 9 367, 309, 179, 124, 79, 74, 9	2% Agarose	CYP2C19P+E1:F2 CYP2C19P+E1:R
IVS1-227	Novel	*Mse*I	37, BSA	AA AG GG	325, 208, 139, 55, 44 325, 252, 208, 139, 55, 44 325, 252, 139, 55	2% Agarose	CYP2C19E2+3:F CYP2C19E2+3:RS1
rs17880188		*HpyCH4*III	37	GG CG CC	293, 89, 41, 29 293, 89, 70, 41, 29 293, 89, 70	15% PAA	CYP2C19E8:F CYP2C19E8:R
rs12248569	*17	*HpyCH4*V	37	CC AC AA	467, 40, 19 486, 467, 40, 19 486, 40	3% Agarose	CYP2C19P+E1:mF*17 CYP2C19P+E1:RS1
rs11568729		*Hha*I	37, BSA	CC CT TT	488, 42 526, 488, 42 526	3% Agarose	CYP2C19P+E1:F3 CYP2C19P+E1:RS1
rs17882687	*15	*BsaB*I	60	AA AC CC	264, 22 286, 264, 22 286	3% Agarose	CYP2C19P+E1:mF*15 CYP2C19P+E1:R
rs12769205		*Bfa*I	37	AA AG GG	222, 84, 72 306, 222, 84, 72 306, 72	2% Agarose	CYP2C19E2+3:FS1 CYP2C19E2+3:R
rs17884712	*9	*HpyCH4*III	37	GG AG AA	279, 98 337, 279, 98 337	3% Agarose	CYP2C19E2+3:FS1 CYP2C19E2+3:R
rs4244285	*2	*Sma*I	25	GG AG AA	410, 113 523, 410, 113 523	2% Agarose	CYP2C19E5:F CYP2C19E5:R
V347I	**28*	*HpyCH4*IV	37	GG AG AA	305, 134, 123 305, 257, 134, 123 305, 257	2% Agarose	CYP2C19E7:F CYP2C19E7:R
rs4451645		*Nla*III	37, BSA	AA AG GG	204, 72, 6 204, 72, 54, 28, 6 204, 54, 18, 6	15% PAA	CYP2C19E9:mF CYP2C19E9:R
rs4986893	*3	*BsaJ*I	60	GG AG AA	244, 124, 104 244, 228, 124, 104 244, 228	3% Agarose	CYP2C19E4:F CYP2C19E4:R
rs6413438	*10	*Bsl*I	55	CC CT TT	108, 24, 10 132, 108, 24, 10 132, 10	15% PAA	CYP2C19E5:F CYP2C19E5:mR*10

Supplementary Table C: Identified variants in the *CYP2C19* gene in the Xhosa population

Allele	Variant	Region	Effect on protein	Effect on enzyme activity	Frequency in Xhosa	Xhosa Individuals genotyped	Frequency in CMA	CMA Individuals genotyped
*CYP2C19*15*	**- 2030C>T**	5'		Unknown	0.11	50	0.08	74
	- 2020C>A	5'		Unknown	0.11	50	0.08	74
	-1439T>C rs17878739	5'			0.11	50		
	55A>C rs17882687	Exon 1	I > L	Unknown	0.09	100	0.08	66
*CYP2C19*9*	**431G>A rs17884712**	Exon 3	R > H	Decrease in vitro	0.09	100	0.04	75
	IVS1+11T>C rs17882201	Intron 1			0.07	49		
*CYP2C19*2*	**- 98T>C rs4986894**	5'			0.17	15		
	IVS1-231G>A rs7916649	Intron 1			0.77	15		
	681G>A rs4244285	Exon 5	Splicing defect	Defective	0.21	100	0.17	75
	IVS5+228G>A rs12571421	Intron 5			0.17	15		
	IVS5-51C>G rs4417205	Intron 5			0.20	15		
	IVS6-196T>A rs28399513	Intron 6			0.17	15		
	990C>T rs3758580	Exon 7			0.13	15		
	991A>G rs3758581	Exon 7	I > V	None	1.00	15		
*CYP2C19*17*	**-806C>T rs12248560**	5'		Increase	0.10	100	0.14	74
*CYP2C19*3*	**636G>A rs4986893**	Exon 4	Premature stop codon	Defective	0.00	100	0.07	75
*CYP2C19*28*	**1120G>A**	Exon 7	V > I	Unknown	0.01	100	0.01	75
*CYP2C19*27*	**-1041G>A rs7902257**	5'		Decrease in expression *in vitro*	0.33	97	0.08	74
	-1418C>T rs3814637	5'			0.22	50		
	-783C>T rs11568729	5'		Unknown	0.08	100		
	99C>T rs17885098	Exon 1	None	None	0.83	15		
	IVS1+20C>T rs17881883	Intron 1			0.03	15		
	IVS1-340T>C rs17884832	Intron 1			0.10	15		
	IVS1-227G>A	Intron 1			0.05	50		
	IVS2-48C>T	Intron 2			0.03	15		
	IVS2-23A>G rs12769205	Intron 2			0.22	100		
	IVS4-40T>C rs57752480	Intron 4			0.07	15		
	IVS7-106T>C rs4917623	Intron 7	Unknown	Unknown	0.03	15		
	IVS8+122G>C rs17880188	Intron 8			0.03	50		
	IVS8-17A>G rs4451645	Intron 8			0.03	98		
	IVS8-119C>T rs12268020	Intron 8			0.17	15		