UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Prevalence of genetic polymorphisms associated with anti-cancer drug efficacy and toxicity in the South African population

by

Bianca Sansom

Thesis submitted in fulfilment of the requirements for the degree

Philosphiae Doctor

in the Department of

Pharmacology

in the

Faculty of Health Sciences

at the

University of Pretoria

Supervisor
Prof Vanessa Steenkamp

Co-supervisor
Prof Elizabeth Jansen van Rensburg

December 2021
© University of Pretoria

Declaration of originality

I, Bianca Sansom,

Student number: 04417577

Subject of the work: Prevalence of genetic polymorphisms associated with anti-cancer drug efficacy and toxicity in the South African population

Declaration

1. I understand what plagiarism entails and am aware of the University's policy in this regard.

2. I declare that this project (dissertation) is my own, original work. Where someone else's work was used (whether from a printed source, the internet or any other source) due acknowledgement was given and reference was made according to departmental requirements.

3. I did not submit this work previously (except for published original articles which have resulted from findings in this study before the submission of this completed thesis), nor has it been submitted to any other university for any degree of examination, either by myself or any other party.

4. I did not make use of another student's previous work and submitted it as my own.

5. I did not allow and will not allow anyone to copy my work with the intention of presenting it as his or her own work.

Signature:

Date: 23rd of February 2022

Abstract

Cancer is a leading cause of morbidity and mortality in both industrialized and developing nations. The true incidence of cancer in unknown in many developing countries, where the burden of the human immunodeficiency virus and tuberculosis take priority. Lung cancer and female breast cancer have the highest number of new cases globally, followed by colorectal cancer, prostate cancer, and stomach cancer. Treatment failure is not uncommon in these types of cancer and are often life threatening. To improve treatment outcome, focus has shifted to incorporate precision medicine. This involves pharmacogenomics which aims to understand how genetic variations influence drug efficacy and toxicity, and how these individual genetic variations affect the manner in which an individual will respond to a specific drug regime or dosage. There are some population groups for which there is limited genetic variation data sets available, such as sub-Saharan African populations. The aim of the study was to establish the prevalence of several single nucleotide polymorphisms (SNPs) that have been highlighted to not only affect the efficacy and toxicity of anti-cancer drugs (but other drugs as well) within the South African population of Bantu speakers, bridging a gap in knowledge which could contribute towards improved treatment strategies for cancer treatment.

The study protocol was approved by the Ethics Committee of the Faculty of Health Sciences of the University of Pretoria under protocol number 3/2020. A total of 764 DNA samples from four previously recruited cohorts of self-identified black South African Bantu speakers were assessed. The four cohorts included blood bank donors for the study of gene polymorphisms in the general population (Ethics protocol # 73/2006), women with breast cancer for the study of breast cancer genes (Ethics protocol #KCT265), men with prostate cancer for the study of prostate cancer genes (43/2010), and lastly male Tshivenda individuals who were previously enrolled as controls in a study investigating epigenomic changes due to pesticide exposure in a malaria area (Ethics protocol # 43/2003). Sixty genetic polymorphisms, consisting of a single target gene, 13 transport genes and 46 metabolisms genes were investigated. SNPs comprised of missense variants (60%), intron variants (13.3%) and synonymous variants (8.3%). The TaqMan® OpenArray™ genotyping platform was used to genotype the study cohort. The OpenArray™ platform was validated using whole

genome sequencing. Genotype and allele frequencies for the study population were calculated and compared to data extracted from the 1000 Genomes Project on populations of African and European ancestry to determine the presence of interpopulation variation. Statistical significance ($P < 0.05$) was determined using both the Chi-square and Fisher exact test. Principle component analysis (PCA) was performed to visualise the variation both between the populations and within the study population.

Six polymorphisms showed 100% amplification across all the samples analysed; rs1272155 (*CYP2B6*), rs11572103 (*CYP2C8*), rs35599367 (*CYP3A4*), rs55886062 (*DPYD*), rs714368 (*SLC22A16*) and rs1902023 (*UGT2B15*). The *CYP2D6* locus (rs1065852, rs3892097, rs35742686 and rs28371706) had the highest no amplification call rate of all the SNPs investigated. Five variants were fixed within the population, with only homozygotes for the ancestral allele being detected in the study population namely, rs2032582 - C *(ABCB1),* rs12721655 - A (*CYP2B6*), rs7900194 – G (*CYP2C9*), rs55886062 - A and rs67376798 - T (*DPYD*). Six polymorphisms, rs1065852, rs3892097, rs35742686 and rs28371706 from *CYP2D6* family*,* rs11615 (*ERCC1*) and rs1042028 (*SULT1A1*) were found to deviate from HWE ($P < 0.05$), likely because of the non-specificity of the probes rather than population stratification.

The PCA analysis confirmed the presence of both inter- and intra-population diversity, highlighting the importance and need for more large-scale population-based studies of understudied populations. The mechanism of action of the drug and the role that the gene or subsequent enzyme plays in the metabolism of the drug, is the determinant of the risk variant. From the study several of the SNPs were highlighted as important polymorphisms for screening purposes prior to therapy initiation, which is in agreement with previous literature; *CBR3* (rs1056892), *CYP19A1* (rs4646), *CYP2B6*16* (rs28399499), *CYP2B6*6* (rs3745274), *CYP2C19*17* (rs12248560), *CYP2C19*27* (rs7902257), *CYP2C8*2* (rs11572103), *CYP2D6*17* (rs2837170), *CYP3A4*1* (rs4986907), *CYP3A4*1B* (rs2740574), *CYP3A5*3* (rs776746), *CYP3A5*6* (rs10264272), *CYP3A5*7* (rs41303343), *DPYD* (rs115232898), *SLCO1B3* (rs4149117 and rs7311358), *UGT1A6* (rs17863783) and *XRCC1* (rs25487).

Future research should include genomic capacity development in the African continent. Policy developers need to enforce pharmacovigilance reporting, which will enable large pharmaceutical companies to assist in research and funding of projects that will investigate the underlying cause of adverse drug reactions in specific population groups within Africa.

Keywords: Adverse drug reactions, breast cancer, cancer treatment, pharmacodynamics, pharmacokinetics, pharmacogenetics, prostate cancer and single nucleotide polymorphisms**.**

Acknowledgements

In loving memory of my grandmother and grandfather, Amy Sansom, who was at the fore front of my early development and learning, and Melville Reynierse, who remains my motivation to continue my research in the field of cancer genetics.

Although I am personally proud of this accomplishment, I know that I do not submit this thesis alone, but instead with an army of support behind me. There is a long list of people I need to thank for this achievement. Firstly, to my husband (Andre Pieterse), who has never faltered in his support. To my parents, thank you for the genes of determination and drive, that continue to push me towards my goals. Thank you to Uncle Shaun, for always having the faith and belief in me, it is your pride in me which guides me to follow my dreams.

No academic accomplishment is complete without the support, guidance, and late nights that supervisors sacrifice, to aid their student in accomplishing their goals. I think, the distance between us the past two years has been a new learning curve for both student and mentor. Prof Steenkamp, you are a pillar of strength, and a leader that many look up to and admire. The year 2020, has brought numerous challenges but never once, have you stopped or given up. Prof Jansen van Rensburg, your knowledge and insight into the world of genetics first caught my attention in the third-year human genetics course you taught. I remember our late night at the lab, surfing the web for single nucleotide variations, it's that commitment that has pushed me to complete this degree. Prof Michael Kidd, from the Department of Statistics, University of Stellenbosch, was instrumental in the assistance and guidance provided for the statistical analysis of the project. Understanding population genetics, and the associated rules is not an easy task. I thank you for your time and patience with myself and the study.

I am a strong believer is grasping all opportunities firmly with both hands, jumping headfirst and just going for it, giving it your all. This fearless temperament has allowed me to establish a thriving career, it has also presented me with funding opportunities and expansion on my scientific connections. I would thus like to thank Rizwana Mia, from the South African Medical Research Council for the assistance in securing

reagents for the project. I would also like to thank, Judith Hornby and Aubrey Shoko from the Centre for Proteomics and Genomic Research (CPGR), for allowing me the opportunity to conduct my lab work at their facility. Finally, to the University of Pretoria, for the financial support in the form of the postgraduate bursary that I received over the past three years which has enabled me to complete my PhD degree.

I end this acknowledgement with a quote from Regina Brett, "*No matter how you feel, get up, dress up, show up, never give up*". Even, during all the craziness and the COVID-19 pandemic, the words "*never give up*" resonate with me wherever I go and whatever new challenge I undertake.

List of abbreviations

5-FU	Fluorouracil
ABC	ATP-binding cassette
ACS	American Cancer Society
ADME	Absorption, distribution, metabolism and excretion
ADR/s	Adverse drug reaction/s
ADT	Androgen deprivation therapy
AFR	Africa
AI/s	Aromatase Inhibitor/s
AS	Activity score
ASR	Age standardised incidence rate
BER	Base excision repair
BLAST	NCBI basic local alignment search tool
Bp	Base pair
BrCa	Breast cancer
CA	Correspondence analysis
CPIC	Clinical Pharmacogenetics Implementation Consortium
CUP	Cancer of unknown primary
CYP	Cytochrome P450
DF	Degrees of freedom
DHEA	Dehydroepiandrosterone
DNA	Deoxyribonucleic acid
DPD	Dihydropyrimidine dehydrogenase
DPWG	Dutch Pharmacogenetics Working Group
DSB	Double-strand break repair
EM	Extensive metaboliser
EMA	European Medicines Agency
ER	Estrogen receptor
EU	European
FAC	5-FU, doxorubicin and cyclophosphamide chemotherapy protocol
FDA	US Food and Drug Administration
FdUMP	5-Fluoro-2'-deoxyuridine 5'-monophosphate
FOLFOX	Fluorouracil, leucovorin, oxaliplatin

FUTP	5-Fluorouridine triphosphate
gnomAD	Genome Aggregation Database
GST	Glutathione S-transferase
GWAS	Genome wide association studies
HAAD	High Coverage African Dataset
HIV	Human immunodeficiency virus
HWD	Hardy-Weinberg deviation
HWE	Hardy-Weinberg equilibrium
IM	Intermediate metaboliser
LMICs	Low- to middle- income countries
MADCaP	Men of African descent carcinoma of the prostate
MAF	Minor allele frequency
mCRPC	Metastatic castration-resistant prostate cancer
MGB	Minor groove binder
MMR	Mismatch repair
MTX	Methotrexate
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
NCR	National Cancer Register
NER	Nucleotide excision repair
NHLS	National Health Laboratory Services
NM	Normal metaboliser
NTC	Non-template control
PCR	Polymerase chain reaction
PD	Pharmacodynamics
PK	Pharmacokinetics
PM	Poor metaboliser
PR	Progesterone receptor
PrCa	Prostate cancer
PV	Pharmacovigilance
QC	Quality control
RNA	Ribonucleic acid
RT	Radiation therapy
SA	South Africa

SAHPRA	South African Health Products Regulatory Authority
SAMRC	South African Medical Research Council
SDS	Sodium dodecyl sulfate
SERMs	Selective estrogen receptor modulators
SLC	Solute-carrier transporter
SNP	Single nucleotide polymorphisms
SNV	Single nucleotide variation
SSA	Sub-Saharan Africa
SULT	Sulfotransferase
TAM	Tamoxifen
TB	Tuberculosis
TS	Thymidylate synthase
UK	United Kingdom
UM	Ultra-rapid metaboliser
USA	United States of America
UTP	Uridine triphosphate
WGS	Whole genome sequencing
WHO	World Health Organization

List of genes

ABCB1   *ATP binding cassette subfamily B member 1*

ABCC2   *ATP binding cassette subfamily C member 2*

ABCG1   *ATP binding cassette subfamily G member 1*

ABCG2   *ATP binding cassette subfamily G member 2*

BRCA   *Breast Cancer Gene*

COMT   *Catechol-O-methyltransferase*

CBR3   *Carbonyl reductase 1*

CBR3   *Carbonyl reductase 3*

CYP1A1   *Cytochrome P450 family 1 subfamily A member 1*

CYP1A2   *Cytochrome P450 family 1 subfamily A member 2*

CYP2A6   *Cytochrome P450 family 2 subfamily A member 6*

CYP2B6   *Cytochrome P450 family 2 subfamily B member 6*

CYP2C8   *Cytochrome P450 family 2 subfamily C member 8*

CYP2C9   *Cytochrome P450 family 2 subfamily C member 9*

CYP2D6   *Cytochrome P450 family 2 subfamily D member 6*

CYP3A4   *Cytochrome P450 family 3 subfamily A member 4*

CYP3A5   *Cytochrome P450 family 3 subfamily A member 5*

CYP17A1   *Cytochrome P450 family 17 subfamily A member 1*

CYP19A1   *Cytochrome P450 family 19 subfamily A member 1*

DPYD   *Dihydropyrimidine dehydrogenase*

EPHX1   *Epoxide hydrolase 1*

ERCC1   *Excision repair cross-complementing 1*

GSTM1   *Glutathione S-transferase Mu 1*

GSTP1   *Glutathione S-transferase Pi 1*

GSTT1   *Glutathione S-transferase Theta 1*

HSB3B1	*3-Beta-hydroxysteroid dehydrogenase-1*
MTHFR	*Methylenetetrahydrofolate reductase*
OATP	*Organic anion transporting polypeptides*
OAT	*Organic anion transporters*
OCT	*Organic cation transporters*
SLC	*Solute vector carrier*
SLCO1B1	*Solute carrier organic anion transporter family member 1B1*
SLCO1B3	*Solute carrier organic anion transporter family member 1B3*
SULT2A1	*Sulfotransferase 1A1*
TPMT	*Thiopurine methyltransferase*
TS	*Thymidylate synthase*
TSPYL1	*Testis specific Y-encoded-like protein 1*
UGT	*UDP glucuronosyltransferase*
UGT1A6	*UDP glucuronosyltransferase family 1 member A6*
UGT2B15	*UDP glucuronosyltransferase family 2 member B15*
UTP	*Uridine-5'-triphosphate*
XPD	*Xeroderma pigmentosum group D*
XPG	*Xeroderma pigmentosum group G*

Glossary

Adverse drug reactions: An injury incurred as a result of a single dose or prolonged administration of a drug, or combination of drugs. The study of adverse drug reactions is a field of pharmacovigilance.

Precision medicine: Is a medical model, which focuses on tailoring disease prevention and treatment, based on a patient's gene, environment and lifestyle, moving away from the one-drug-fits-all model. The idea is that a diagnostic test would be employed for the selection of the appropriate and optimal therapy based on the genetic makeup and molecular analysis of the patient.

Pharmacogenetics: Is the study of the role of the genome in drug response, thus how genomic variations in an individual affect their response to pharmacotherapy.

Pharmacogenomics: Analyses how the genetic makeup of an individual affects their response to drugs. Assessing the effect that acquired or inherited genetic variation and gene expression patterns have on the drug metabolism of an administered drug.

Pharmacovigilance: Pharmacological science relating to the collection, detection, assessment, monitoring and reporting adverse effects with pharmaceutical products. The goal of pharmacovigilance is ultimately the prevention of adverse effects, which can only be obtained through surveillance records.

List of Figures

List of Tables

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

The National Cancer Institute (NCI) estimated that in 2020, 1.8 million new cases of cancer will be diagnosed in the United States.[1] The most common cancers were expected to be breast cancer, lung and bronchus cancer, prostate cancer, colon and rectum cancer, melanoma of the skin, bladder cancer, non-Hodgkin lymphoma, kidney and renal pelvis cancer, endometrial cancer, leukemia, pancreatic cancer, thyroid cancer, and liver cancer.[1] Prostate, lung, and colorectal cancers were estimated to account for 43% of all cancers diagnosed in men, whereas breast, lung, and colorectal were estimated to account for 50% of cancers diagnosed in women.[1] Updated statistical data are still to be published.

The health burden placed on low- to middle-income countries (LMICs) is significant. The American Cancer Society (ACS) estimated that approximately 65% of the global deaths will occur in LIMCs in 2020.[2,3] Many of these LMICs are dealing with several other life threatening epidemics such as human immunodeficiency virus (HIV) and tuberculosis (TB), in conjugation with the global pandemic of COVID-19. The COVID-19 pandemic has had a direct impact on the morbidity and mortality of global cancer, as many hospitals and clinics have had to shut down, leading to delays in diagnoses and treatment, but also missed preventative screening and treatment.[2]

Treatment inequalities, and limitations to clinical trial access are also of concern for cancer patients in LMICs.[2] It is well known that cancer patient's response to chemotherapy varies greatly. Many patients experience a complete or partial response, while for some the chemotherapy will be ineffective despite the onset of drug toxicities.[4] Advanced cancer treatments are not always available in LMICs, and treatment progress is limited if the incidence of cancer is not a priority in the region.[2]

Toxicities or adverse drug reactions (ADRs) to drug treatment may be severe and potentially life threatening, leading to hospitalization. Various factors influence the odds of a patient experiencing ADRs, which include baseline clinical and demographic characteristics as well as genetic factors. Although there are efficacious regimes

available, the selection of the safest regime is not guided by molecular biomarkers, but rather age, comorbidities and performance status.[4] The World Health Organization (WHO) initiated an International Drug Monitoring Programme in 1968, that monitors and identifies the harm caused by medicines, in order to reduce the risks to patients, and create a pharmacovigilance standards system called Vigibase.[5]

In the United States of America (USA), ADRs account for approximately 7% of all hospital admissions, with an annual estimated cost to the health care system of $6.3 billion (approximately R91.3 billion).[4] Pharmacovigilance (PV) in South Africa is still a relatively new concept. South Africa (SA) has submitted 28 609 reports (27 reports per million per year) since its inception in 1992.[5] This number is small relative to the real occurrence of ADR cases, as the majority of cases go unreported. Furthermore, considering the high number of people living with HIV, TB and other communicable and non-communicable diseases, a much higher number is expected. A meta-analysis study which included published data on all prospective and retrospective studies where ADRs were investigated, found that ADRs were responsible for 6.7% of all hospitalisations.[6] This burden is higher in developing countries such as SA where the use of self-medication, unregistered and adulterated medicines, as well as traditional and herbal therapies is common.[7]

Pharmacogenomics is used to identify genetic variations that contribute to drug response, to reduce ADRs and increase drug efficacy. With the completion of the Human Genome Project and subsequence projects such as the 1000 Genomes Project, the identification of gene variations associated with the pharmacokinetics (PK) and pharmacodynamics (PD) of drugs became possible.[8] Given the clinical implication of these variations, knowledge of their prevalence within the SA population of Bantu speakers would therefore be important from a public health perspective. Population-based studies are required to infer rates and risks for drug inefficacy and/or ADRs in order to guide the potential implementation of pharmacogenetic testing.[8] The 1000 Genomes Project has created great insight into population diversity, however none of the Southern Africa, African population have been included in these studies. Recent studies have highlighted key differences in allele frequencies within the sub-Saharan Africa (SSA) population, suggesting they are not ideal proxy populations of all Africans.[9,10]

CHAPTER 2: LITERATURE REVIEW

2.1 Cancer epidemiology

2.1.1 Global Statistics

Cancer is considered a national priority as it places a financial burden on the public health sector. According to WHO, cancer is the second leading cause of death globally. In 2018 alone 18.1 million new cases were reported and 9.6 million lives lost due to cancer.[11] It is estimated that there are approximately 43.8 million people living with cancer or within 5 years of cancer diagnosis.[12] The ACS estimated that in 2020, 65% of the 10 million global cancer related deaths occurred in LMICs.[2] They further projected that by 2040, there will be 16 million global cancer deaths (as a result of demographic and epidemiological transitions), of which 69% will occur in LMICs.[2]

According to the GLOBOCAN estimates, lung cancer and female breast cancer (BrCa) are responsible for the highest number of new cancer cases reported globally; and in 2018 accounted for 11.6% of the total cancer incidents; followed by colorectal cancer (10.2%), prostate cancer (PrCa; 7.1%) and stomach cancer (5.7%).[12-14] Population growth, ageing and social economic development contribute to the increasing cancer burden.[11] There are several risk factors that contribute to the development of cancer namely; tobacco usage, obesity, poor diet, lack of physical activity and alcohol consumption, infections, hepatitis B, hepatitis C and human papilloma virus as well as inherited genetic defects.[15]

Worldwide, lung cancer is the highest diagnosed cancer in men, accounting for 14.5% of cases, and is also the leading cause of death in men (22%). This is followed by PrCa (13.5%), colorectal cancer (10.9%), liver cancer (10.2%) and stomach cancer (9.5%).[12-14] In women, BrCa is the most commonly diagnosed cancer (24.2%) and the leading cause of death (15%). Lung cancer (13.8%) is ranked second, followed by colorectal cancer (9.5%) and cervical cancer (6.6%).[12-14]

The Asian continent has the highest incidence of cancer for both genders (48.4%), followed by Europe (23.4%), America (21%) and Africa (5.8%). LMICs account for approximately 70% of the global cancer deaths.[12-14] The highest number of reported

new cases of cancer and related deaths in 2018 was in Asia. In 2018 Asia had the highest number of reported cancer deaths (57.3%), followed by Europe (20.3%), Americas (14.4%) and Africa (7.3%) (Figure 1).[13] For Asia and Africa the mortality percentage is greater than their percentage of incidence. This is ascribed to the high frequency of cancer types with poor prognosis and higher mortality rates (lung and colorectal cancer), in addition to limited access to timely diagnosis and treatment.[12-14]
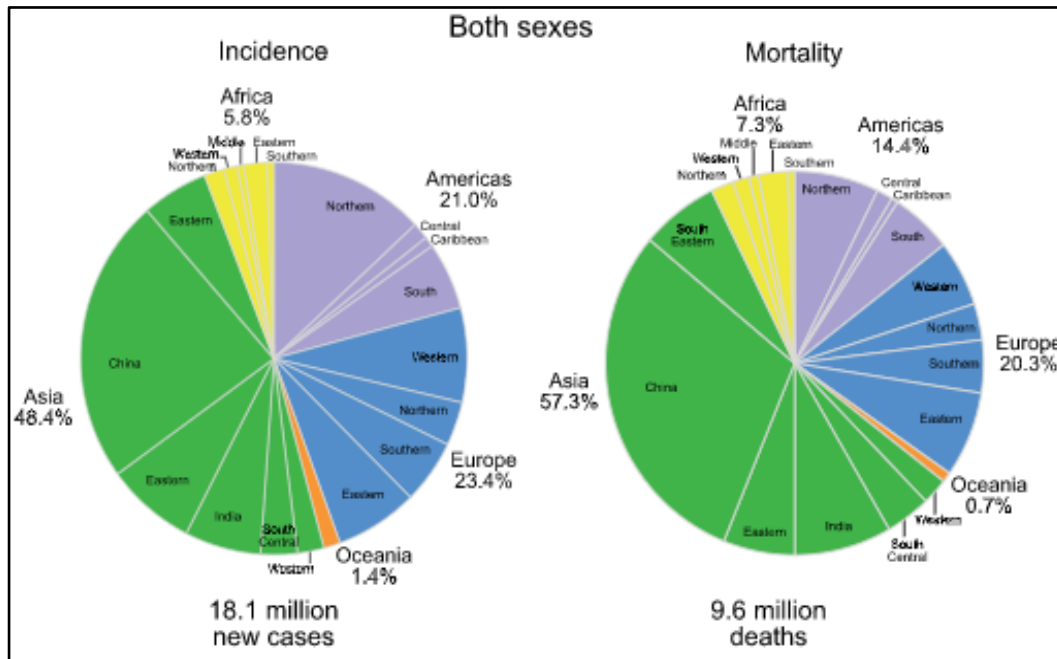
Figure 1. Percentage of new cancer cases and deaths for both sexes globally as in 2018 (Reproduced with permission[13]).

The National Cancer Registry (NCR) is a surveillance unit, with the primary function of collating and reporting on the cancer incidence rates. The latest statistics released by the NCR and the National Health Laboratory Services (NHLS), was in 2016 where PrCa, colorectal cancer and lung cancer were ranked as the three most commonly diagnosed cancers in men in the SA population (Table 1).[15,16] The most commonly diagnosed cancers in women at the same time were BrCa, cervical cancer and cancer of unknown primary origin (CUP) or also referred to as occult primary tumours (Table 2).[15] Although the NCR is the largest repository of cancer data in SA, the data is garnered mostly from pathology-based results, which due to exclusion of other analyses culminates in an under reporting of cases. Cancer reporting has not been a national health priority due to the HIV and TB epidemic. However, Regulation 380 of

the National Health Act 61 of 2003,[17] formally appointed the NCR as South Africa's main cancer surveillance agency, and this Unit made cancer reporting obligatory.[18]

Approximately 80% of people diagnosed with cancer in SSA are already in advanced stages of the disease, with fewer than 10% receiving pain relief treatment, chemotherapy or radiotherapy.[19] Traditionally, cancer was not SSA's biggest healthcare problem. Communicable diseases, such as pneumonia, malaria and HIV/AIDS, have always been the diseases resulting in the most deaths.[20] Furthermore, anti-cancer drugs are many times inaccessible to patients due to the high cost of the drug.[21]

Table 1. The ten most common cancers diagnosed in South African men in 2018.[15]

Type of cancer	Actual number of cases	Age standardised incidence rate (ASR)*	Estimated lifetime risk	Percentage of all cancers
Prostate cancer	8 332	47.48	1:17	21.33
Colorectal cancer	2 065	11.01	1:76	5.29
Lung cancer	1 892	10.23	1:77	4.84
Cancer of unknown primary (CUP)**	1 890	9.7	1:86	4.84
Non-Hodgkin's lymphoma	1 290	5.83	1:160	3.30
Malignant melanoma	1 139	5.88	1:153	2.92
Kaposi sarcoma	1 114	3.95	1:287	2.85
Cancer of the bladder	997	5.61	1:158	2.55
Oesophageal Cancer	964	5.06	1:161	2.47
Cancer of the stomach	803	4.27	1:195	2.06
All cancer	39 060	137.79	1:6	

* ASR: Age standardised incidence rate per 100 000
** Unknown primary (CUP) or occult primary tumour

According to the NCR statistical report, PrCa is the most common cancer in both African and Caucasian males, whereas cervical cancer is the most common cancer in African women, followed by BrCa.[15] In Caucasian females, BrCa is the most common cancer. It has been reported that African women have a higher incidence of BrCa before the age of 40, and are more likely to die of the disease.[15]

Table 2. The ten most common cancers diagnosed in South African women in 2018.[15]

Type of cancer	Actual number of cases	Age standardised incidence rate (ASR)*	Estimated lifetime risk	Percentage of all cancers
Breast cancer	9 548	35.95	1:25	22.58
Cervical cancer	7 327	26.96	1:35	17.33
Cancer of unknown primary (CUP)**	1 964	7.45	1:114	4.64
Colorectal cancer	1 819	7.45	1:127	4.30
Uterine cancer	1 446	5.67	1:135	3.42
Lung cancer	1 088	4.25	1:182	3.80
Non-Hodgkin's lymphoma	1 101	3.94	1:239	2.60
Malignant melanoma	950	3.5	1:262	2.25
Oesophageal cancer	853	3.26	1:255	2.02
Kaposi sarcoma	716	2.26	1:540	1.69
All cancer	42 288	124.42	1:7	

* ASR: Age standardised incidence rate per 100 000
** Unknown primary (CUP) or occult primary tumour

2.1.2 Incidence of prostate cancer in South Africa

The most common PrCa is adenocarcinomas, which accounts for 90% of cases. Other types of PrCa include; sarcomas, small cell carcinomas, neuroendocrine tumours and transitional cell carcinomas.[22] PrCa is the most common cancer in men, with a high prevalence in both African and Caucasian males (Table 3).[23]

Men between the ages of 50 and 79 years, have been reported as at the highest risk of presenting with PrCa.[23] Risk factors associated with PrCa include; age, family history, ethnicity, nationality, diet, obesity, alcohol consumption and use of anabolic steroids.[24]

Table 3. Prostate cancer incidence in South Africa across ethnic groups.[23]

Ethnic group	No of cases	Age standardised incidence rate (ASR)*	Lifetime risk	Percentage of all cancers
All males	8 332	47.48	1:17	21.33
Asian males	213	29.66	1:24	21.82
Black males	3 769	37.34	1:22	29.34
Coloured males	916	54.32	1:15	19.97
White males	3251	77.08	1:10	16.25

* ASR: Age standardised incidence rate per 100 000

2.1.3 Incidence of breast cancer in South Africa

Carcinoma of the breast remains the most prevalent cancer diagnosed in women world-wide and the different types are categorised on where they form (ducts or lobules) and how far they have spread.[25] *In situ* BrCa are non-invasive, with abnormal cells being contained in the lining of the breast.[25] Invasive BrCa is a type where the abnormal cells have spread to the surrounding breast tissue, whereas metastatic BrCa is where the cancer has spread beyond the breast to surrounding organs such as the lungs, bones and brain. Less common types of BrCa include; inflammatory BrCa, Paget disease of the nipple, phyllodes tumour and angiosarcoma.[25]

The incidence of BrCa in SA, based on statistics released by the NCR in 2018, is provided in Table 4. For similar reasons to PrCa, BrCa is under diagnosed. Women between the ages of 40 and 79, have the highest reported BrCa diagnosis across all ethnic groups.[26] African women between the ages of 30 and 39, have a significantly higher frequency of BrCa diagnosis compared to other ethnic groups.[26]

2.2   Cancer regimens

There are several anti-cancer drugs available, which are administered either as a single drug or a combination of drugs depending on the diagnosis. Certain factors need to be considered when determining a treatment plan for any type of cancer.[27] These include the rate of growth, stage or grade of the cancer, history of previous cancer episodes and overall general health of the patient.[28] The ultimate goal is to cure cancer, however, if this not possible, treatment to shrink the cancer or slow the

progression to enable the patient to live as long as possible under symptom free conditions is considered.[28]

Table 4. Breast cancer incidence per ethnic group in the South African population.[26]

Ethnic group	Number of cases	Age standardised incidence rate (ASR)*	Lifetime risk	Percentage of all cancer
All females	9 548	35.95	1:25	22.58
Asian females	458	50.87	1:18	36.82
Black females	4 159	22.57	1:41	21.41
Coloured females	1 320	48.81	1:19	28.53
White females	3 330	84.58	1:11	20.57

* ASR: Age standardised incidence rate per 100 000

2.2.1 Prostate cancer

The most common treatment options for PrCa are surgery, radiotherapy and hormone therapy. Chemotherapy is infrequently used as a treatment option.[23] Typically, men presenting with early-stage disease or cancer confined to the prostate will undergo radical prostatectomy. Radiation therapy (RT) is used as an initial treatment of PrCa that is localized, and in men that present with advanced or recurrent PrCa. Radiation colitis is a side effect of RT, affecting 20% of patients with PrCa.[23] Androgen deprivation therapy (ADT) forms part of the standard care for advanced metastatic PrCa, whereas chemotherapy is used in cases where metastatic PrCa is diagnosed, despite initiation of hormone therapy, also referred to as metastatic castration-resistant prostate cancer (mCRPC).[23] ADT and its effect on gene-drug interactions are discussed later in the chapter.

2.2.2 Breast cancer

There are several histologic types of BrCa and the treatment options depend on the type and staging of the cancer. *In situ* BrCa is usually treated with local therapies such as surgery and/or radiation.[29] Chemoprevention strategies with selective oestrogen receptor modulators (SERMs) such as tamoxifen (TAM) or aromatase inhibitors (AI), are offered in these instances. Invasive BrCa is treated with local and systemic therapies, depending on staging.[29] Systemic therapies, include, chemotherapy,

hormonal therapy and targeted therapy.[29] SERMs and their effect on gene-drug interactions is discussed later in the chapter.

2.3 Drug efficacy, failure, and toxicity

The PK of anticancer drugs are highly variable and it is this variability coupled to the drugs narrow therapeutic index that makes pharmacokinetic optimisation difficult.[30] An individual's unique genetic makeup can influence disease development and as well as drug response.[31] These inter-individual variations can either be inherited or acquired and pose a problem when dose optimization is required.[31] In many instances individuals do not respond fully to first-line treatments, and dosage or treatment regime adjustments may be required in order for a patient to benefit fully from the treatment.[30] Cancer drugs have been shown to only be effective in 25% - 60% of patients.[31] Two individuals taking the same drug at the same dose can respond differently, this implies that an effective drug dose in one individual may prove to be lethal or result in therapeutic failure in another.[32] Drugs with narrow therapeutic indexes and serious side effects require continuous monitoring. Drug-drug and drug-disease interaction can also exacerbate efficacy of treatment.[31]

Factors that have been associated with varied drug response include physiological variation (weight, height and age) and genetic variation (in genes affecting both pharmacokinetic and pharmacodynamics responses of drugs).[31] These factors can work alone or in combination to influence drug response.[31] The factors that may affect drug response are provided in Figure 2. Mechanisms that enable drug resistance include; drug inactivation, drug target alterations, drug efflux, deoxyribonucleic acid (DNA) damage repair and inhibition of cell death.[33]
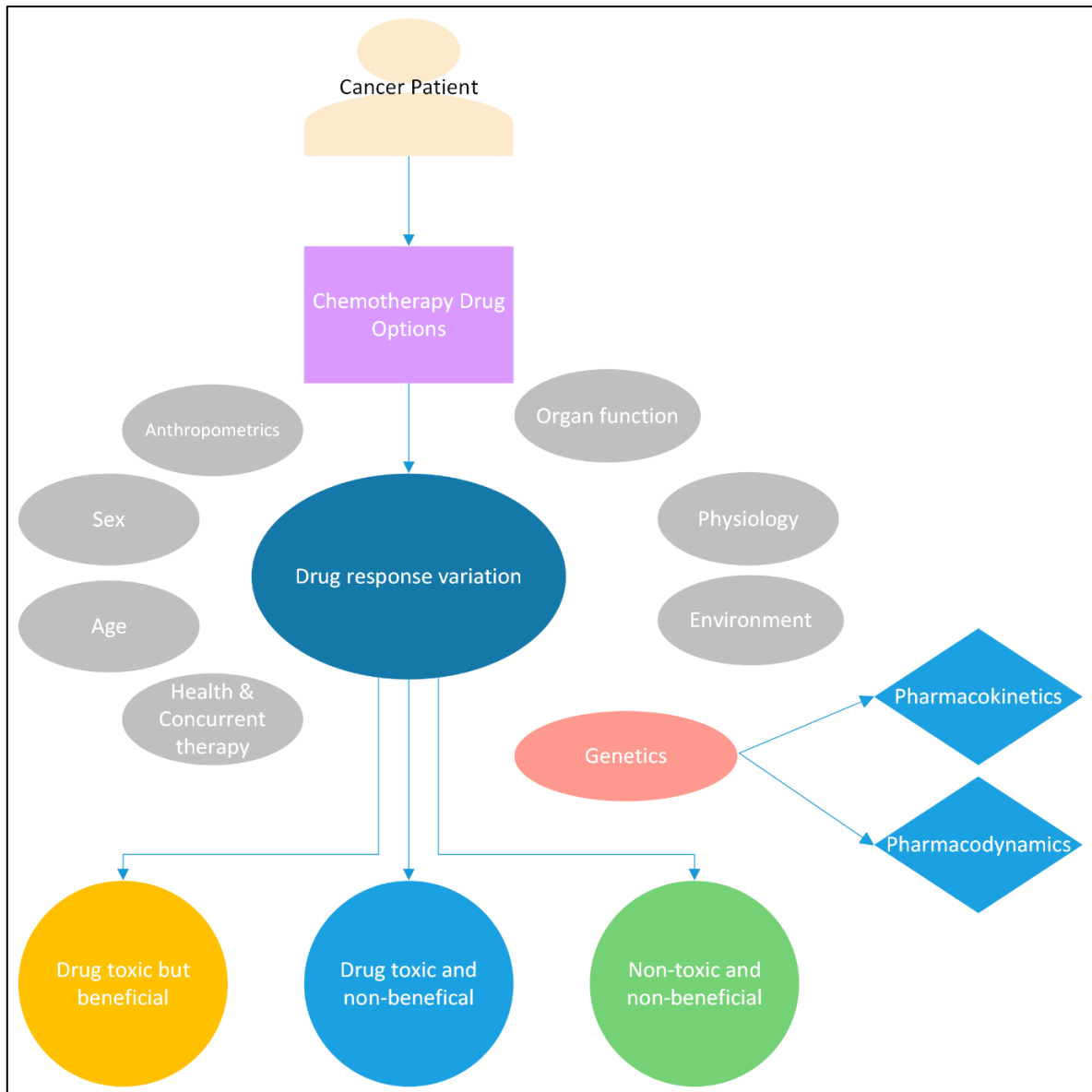
Figure 2. Factors affecting inter-individual variation in drug response.[34]

2.4 Role of pharmacokinetics and pharmacodynamics

Pharmacology is the study of drug or medication action, which exerts a biochemical or physiological effect on the cell, tissue, organ, or organism. This involves two main areas PD and PK.[35] PD focuses on the physiological effects of a drug on the body, while PK focuses on the effect the body has on a drug. The latter is influenced by four criteria; absorption, distribution, metabolism and excretion (ADME; Figure 3).[35]

ADME
Refers to the absorption, distribution, metabolism and excretion of a compound. For a chemical compound to become a marketable drug, the compound is required to have favourableproperties in addition to efficecy.

DISTRIBUTION
The movement of the compound around the body and its propensity to accumulate in certain tissues and organs.

METABOLISM
The body's ability to break down the compound, normal by the liver. Issues of concern during this phase is drug-drug interaction, and the effects of the metabolites.

ABSORPTION
A compound's ability to pass through various barriers in the body such as the intestiinal lining, nasal lining , lungs and skin

EXCRETION
The rate and manner in which the compound exits the body

Figure 3. Overview of the absorption, distribution, metabolism and excretion system of the body**.**

PD places emphasis on dose-response relationships, specifically the relationship between the drugs' concentration in the system and its intensity and time course of effect.[36] The pharmacologic response depends on the drug binding to its target and the level of binding. The response of receptors may be affected by the presence of drugs competing for the same drug target site, the functional state of the receptor or pathophysiological factors due to disease genetic mutations or ageing.[36]

PK refers to the movement of drugs into, through and out of the body. Speed of onset, the intensity and the duration of the response is usually dependent on the following parameters:

i) The rate and extent of uptake of the drug from its site of administration. Factors such as compound solubility, gastric emptying time, intestinal transit, chemical instability in the stomach and permeability of intestinal wall, affect the extent to which a drug is absorbed.

ii) The rate and extent of distribution of the drug to the site of action. Compounds need to be carried to its effector site, most often via the bloodstream. Factors

influencing drug distribution include regional blood flow rates, molecular size, polarity, binding of serum to protein and forming a complex.

iii) The rate and extent to which the compound in broken down.[37] The majority of small-molecule metabolism is carried out by the liver by redox enzymes, termed cytochrome P450 enzymes. The parent compound is converted to a new compound, a metabolite. Metabolites can be more pharmacologically active, than the parent drug should the parent drug be a prodrug.

iv) The rate of elimination of the drug from the body. Compounds and their metabolites are removed from the body via excretion, usually through the kidneys or in the faeces.[37]

2.5   Precision medicine

"Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes – and to give all of us access to the personalized information we need to keep ourselves and our families healthier." - President Barack Obama, during his State of the Union Address, on 20 January 2015. In his address, President Obama expressed a strong conviction that science has the ability to improve health and health care, through precision medicine.[38]

The terms precision personalised or individualised medicine are often used interchangeably, however the preferred term is precision medicine and is driven by new diagnostics and therapeutics. Stratified medicine, however, might be a more accurate term, according to the Priority Medicines for Europe and the World Report of 2013, which focuses on biomarker-based stratification of patient populations – "the right drug to the right patient" and the minimization of adverse side effects.[39] The concept of precision medicine, is centred around the prevention and treatment of an individual by taking into account their variability, even if they present clinically similar, by focusing on genetics, biomarkers, phenotypes or psychosocial characteristics.[38,40] The goal is to improve clinical outcomes of individual patients and minimise the side effect/s in those individuals less likely to respond to a particular treatment.

The concept of precision medicine is not new; blood typing has been around for more than a century to guide blood transfusions; management of infectious disease involves

the identification of the causative organism which is followed by the selection of an effective antimicrobial agent; while bacterial infection is concerned about the drug sensitivity of the causative organism to determine antibiotic treatment. However, new advances in science and technology have narrowed this gap and will continue to do so.[38,40]

Furthermore, advances in technology have had a major impact on genomic studies as it has led to the development of large-scale genome databases and computational tools for analysing large data sets. However, rigorous testing is still required to build the evidence base needed to guide clinical practice and for the establishment of guidelines.

Cancer's treatment is no different. When a person is diagnosed with cancer, the individual is usually placed on the same treatment as another patient without taking the staging of cancer into consideration. Such treatments may be toxic and have severe side effects, as individuals respond differently to treatment.[41] The reason being that individuals form a unique genomic signature.[2] Using the genetic changes in a patient's germline and tumour to determine unique treatment is known as precision medicine.[41] The five rights of medication in the standard for safe medication practices are: the right patient, the right drug, the right dose, the right route and the right time. However, errors including lethal mistakes still occur.[42]

2.6 Pharmacogenomics

Pharmacogenetics is the study of inherited genetic differences in terms of drug metabolic pathways, which can affect an individual's response to drugs and is associated with effectiveness and adverse effects.[43] Single nucleotide polymorphisms (SNPs) in important drug metabolising enzymes affect the transportation, metabolism and receptor binging ability of drugs, further influencing the PK and PD properties of medications.[44] Pharmacogenetics can play a significant role in cancer in terms of prognosis, treatment response and potential of adverse side effects.[45] Pharmacogenomics is one of the fundamental elements of precision medicine and focuses on understanding the difference between therapeutic response between individuals through the utilization of genomic, proteomic, transcriptomic and metabolic

based information.[31,46] Treatment of patients should be aimed at a multifaceted 'omics'.[47] The genomics era has enhanced our knowledge of cancer biology, highlighting the heterogeneity between populations and tumour histology at a whole-genome level.[46,47]

Drug response is primarily determined by the PK and PD nature of the prescribed drug/s, which is either directly or indirectly affected by genetic polymorphisms in phase-1 drug metabolising enzymes and/or transporters.[11] Allele frequencies in these genes differ between populations/groups. The clinical information and molecular data are integrated in order to better understand the biological basis of the disease and select medication that would improve patient outcomes. Whole genome sequencing has allowed scientists to identify specific genetic variation that may be responsible for varied drug response. Whole-genome SNP profiling, haplotyping, multigene analysis and gene expression studies using biochips or microarrays are currently being used to study individual responses to drugs at various levels and could facilitate drug discovery and development by improving the identification of disease and drug targets.[48] Population studies will improve the dose recommendation to identify populations that will benefit from a normal dose and those that will require a dosage adjustment.[36]

2.7   Single nucleotide changes

SNPs are the most common form of genetic variation. SNPs represents a change in a single nucleotide in the genome that differs between members of a species.[49] The Human Genome Project and subsequent projects have enabled the identification of millions of SNPs.[49] They occur approximately once in every 300 to 1000 base pairs (bp). There are more than 14 million SNPs distributed throughout the human genome.[49] There are a vast number of genetic variants in genes that are involved in the PK and PD of drugs that have been shown to influence individual drug response behaviours.[31,50] A single nucleotide variant (SNV) is a variation in a single nucleotide without any limitation of frequency and may arise in somatic cells, such as those caused by cancer.[50]

SNPs can occur in the coding sequence of genes, non-coding regions of genes or intergenic regions.[50] Not all SNPs within the coding regions affect the amino acid sequence of the protein.[50] There are two types of SNPs that can occur in the coding region namely, synonymous and non-synonymous SNPs. Synonymous SNPs do not affect the protein sequence, while non-synonymous SNPs change the amino acid sequence and the protein (Figure 4).[50] The SNPs that occur in the non-coding regions can affect gene splicing, transcription factor binding, messenger ribonucleic acid (RNA) degradation and RNA sequences.[50] Many of the SNPs are polymorphic, i.e. these are variants that occur in different populations at varying frequencies. Variations in the DNA sequence can affect how humans develop disease and respond to pathogens, chemical, drugs, vaccines and other agents.[50] Polymorphisms in the coding regions of an enzyme for example, can either lead to an increased or decreased enzyme activity, thereby affecting the manner in which the drug is metabolised.[51]
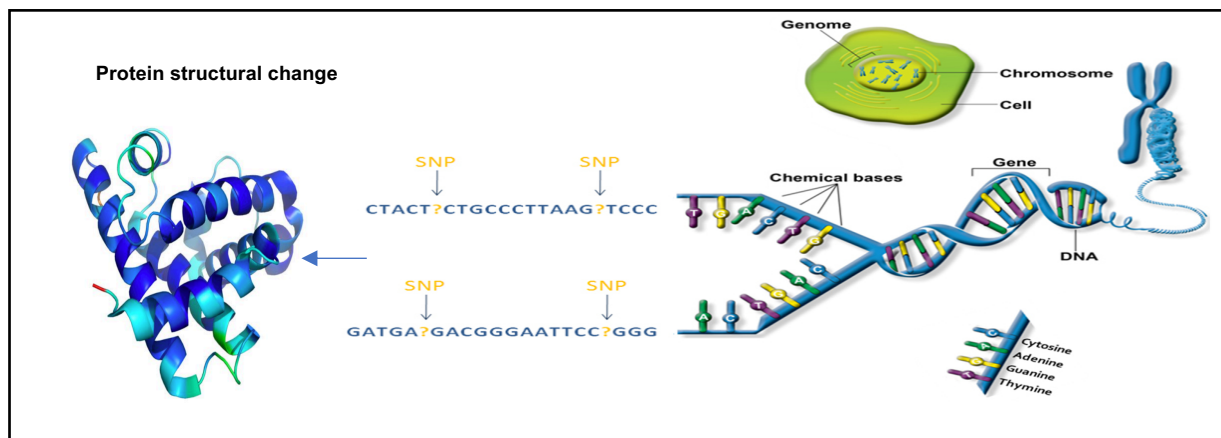
Figure 4. Illustration of DNA, genes and SNPs, indicating the effect that SNPs could have on protein structures when altering the functionality of the protein (Adapted and reproduced with permission[52]).

2.7.1 Penetrance

When considering phenotype, the penetrance of the gene also needs to be considered. The definition of penetrance in genetics refers to the proportion of individuals in a population that carry a particular gene, that also express an associated trait or phenotype. In simple terms, penetrance is a measurement of the relationship between a genotype and phenotype.[53] In instances of complete penetrance, all of the individuals in the population that carry a certain genotype express the corresponding

phenotype. In case of incomplete penetrance, less than 100% of individuals express the corresponding phenotype. For accurate measurement of penetrance, genotype and phenotype recording is required in large populations. A well-studied penetrance gene is that of Breast cancer gene 1 (*BRCA1) and BRCA2*, which has been associated with an elevated risk of breast (5%) and ovarian cancer (10%) in women. Penetrance is difficult to determine accurately as genetic cofactors and epigenetic factors affect the regulation of expression.[53]

2.7.2 Influence of polymorphisms on genes encoding phase-I drug metabolising enzymes

The identification of an individual's genotypes in key enzymes which are involved in the transportation, metabolism and clearance, is a strong determinant of therapeutic efficacy or toxicity.[54] Individuals with no or reduced rates of enzyme activity compared to normal enzyme activity, have higher circulating drug concentrations, providing for a greater chance of drug toxicity developing when administered at the standard dose.[54] On the other hand, individuals with a higher or faster enzyme activity, have a lower circulating drug concentration and experience a sub-therapeutic response at a standard dose. For prodrugs the situation is reversed. Prodrugs require enzymatic activation, to convert the inactive metabolite into an active metabolite.[55] Poor metabolisers (PM) are unable to produce sufficient quantities of bioactive drugs to produce the desired therapeutic effect, while rapid metabolisers produce too much of the bioactive drug.[55]

Analysis of the genetic makeup for metabolic genes is a means to assess rates of metabolic capacity, through the assessment of germ-line polymorphisms.[54] The classical phenotype classification scheme for a specific drug is to compare individuals with variant activities against those who were considered normal or wildtype "extensive metabolisers" (EM) or normal metaboliser (NM).[54] PMs are associated with the presence of null genotypes, intermediate metabolisers (IM) with reduced metabolism genotypes, and ultra-rapid metabolisers (UM) with gene duplications (Figure 5).[54] The activity score (AS) has been broadly accepted to translate genotype to phenotype. In short, each allele is assigned a value of 0, 0.5 or 1, indicative of; no function, decreased or normal function. The sum of the values provides the AS of the genotype.

The scoring system of *CYP2D6* has however recently been amended to ensure standardization, the scoring is as followings: AS = 0 is PM, AS between 0.25 and 1 is IM, AS between 1.25 to 2.25 is NM, AS > 2.25 is UM.[56] The phenotype classification determines the pharmacotherapy adjustment, and possibly a dosage adjustment.
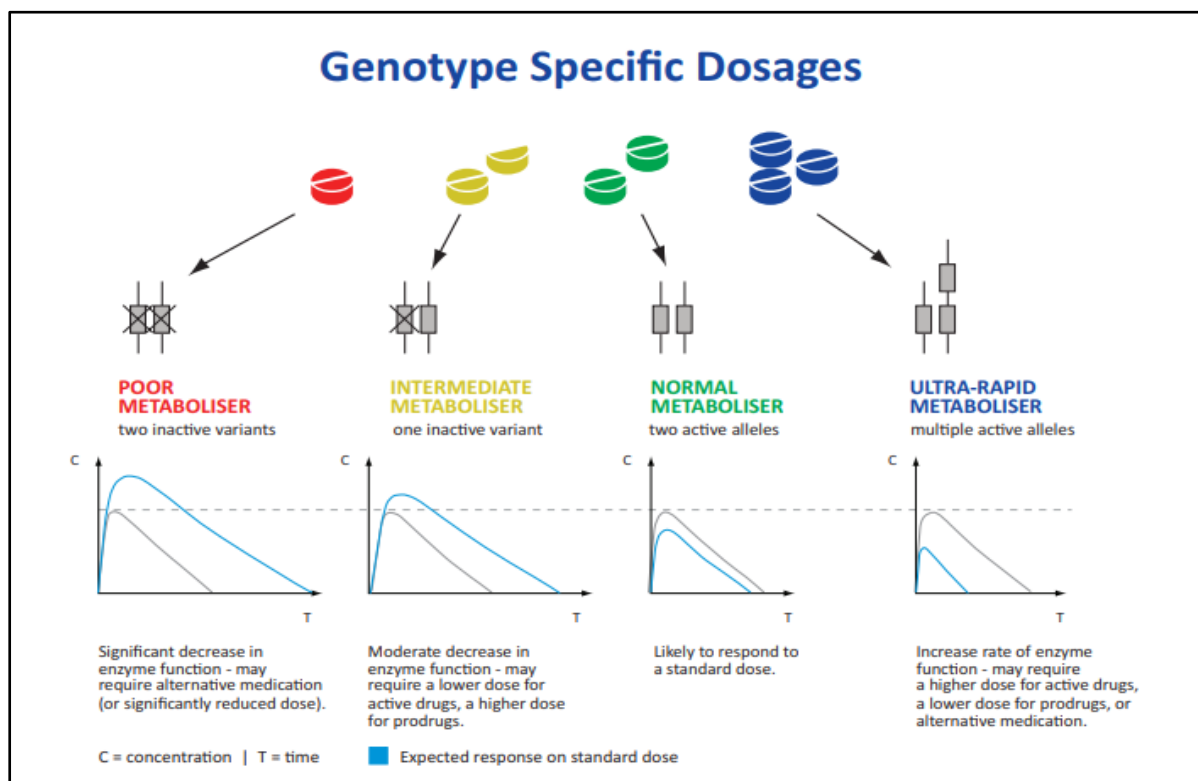
Figure 5. Genetic variations and corresponding metaboliser classification (Reproduced with permission from GeneWay™ Laboratories).

Cytochrome P450 (CYP) represents a large and diverse group of heme-containing enzymes that are involved in the oxidative metabolism of structurally diverse molecules such as drugs, chemicals and fatty acids.[57,58] Genes encoding CYP enzymes and the enzymes themselves are designated with the root symbol CYP for the superfamily, followed by a number which indicates the gene family, thereafter a capital letter, which indicates the subfamily and lastly another numeral, which indicates the individual gene, e.g. *CYP2D6*. More than 100 *CYP2D6* genetic variants have been described.[57,58] These have been derived from point mutations, duplication, insertions or deletions of single or multiple nucleotides and even whole-gene deletions.[57] The clinical implication of genetic polymorphisms of various *CYP* genes is important. The mechanism of action thereof is provided in Table 5. The following African countries do

not have *CYP* data available; Burundi, Cape Verde, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Lesotho, Mauritania, Mauritius, Niger, Somalia, Swaziland, Togo, Western Sahara and Zambia.[59]

Table 5. Genetic polymorphisms of CYP P450s and their effects on the mechanism of action on the specific drug provided as an example.

CYP gene	Drug example	Mechanism of action
CYP2D6	Tamoxifen	*CYP2D6* hydrolates aromatic rings or aryl-alkyl amines containing protonated nitrogen. *CYP2D6* is responsible for the catalysation of tamoxifen to endoxifen and 4-hydroxytamoxifen. The levels of endoxifen vary depending on *CYP2D6* activity.[31,54]
CYP2C9	Warfarin	The enzymes oxidize neutral and acidic amphipathic drugs with a hydrophobic regions.[54]
CYP2C19	Clopidogrel, PPIs*	Acts on weakly or strongly basic drugs containing a hydrogen bond donor, or functional groups containing carbon or sulphur, double bonded to oxygen.[54]
CYP3A4/ CYP3A5	Abiraterone, paclitaxel and docetaxel	Involved in the metabolism and excretion of xenobiotics in the body. Contains monooxygenases properties which catalyses many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.[60,61]
CYP19A1	Aromatase inhibitors	This gene provides instructions for making the enzyme aromatic. Inhibitors of this gene, could halt or alter the production of the aromatic enzyme.[60,61]
CYP1A2	Aromatase inhibitors, caffeine and melatonin	Involved in the metabolism and excretion of xenobiotics in the body. Contains monooxygenases properties which catalyses many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. [60,61]

* PPI: Proton pump inhibitors

2.7.3 Influence of polymorphisms on genes encoding phase-II drug metabolising enzymes

Glutathione S-transferases (*GSTs*) are an important phase II metabolic enzyme family. These proteins are responsible for catalysing electrophilic material and combining substrates with glutathione.[62] They are capable of combining lipotropic cytotoxic agents, which enhances their solubility thereby promoting drug excretion and reducing their effect in the body.[63] Tumour cells are capable of expressing *GSTs* to protect

themselves against the effects of chemotherapy drugs, leading to tumour drug resistance.[62] *GSTM1, GSTT1* and *GSTP1* show a high polymorphic distribution, which modifies the way in which corresponding enzymes detoxify heterogeneous substances.[62] The diminished enzyme function of *GST* genes, results in a decreased deactivation function of the drugs enzyme, which may result in the improvement of the efficacy of the administered drug.[62]

2.7.4 Effect of polymorphisms on genes encoding drug transporters

A drug can either have a favourable or toxic effect. The nature and extent thereof is largely dependent on the ADME of the drug.[32] Drug membrane transporter proteins control the movement of all drugs, as well as their metabolites (active or inactive), across the cell membrane.[62] The presence of polymorphisms in drug transporter genes can alter the ADME rate, and thus ultimately treatment; be it that resistance, toxicity or efficacy of the drug is achieved.[32] The ATP-binding cassette (ABC) and solute-carrier transporters (SLC) are membrane-bound transport proteins that protect the cells and tissues from environmental toxins.[62] Mutations of ABC proteins may lead to multidrug resistance, by way of decreasing effective intracellular concentrations of the drugs.[62]

The *ABCB1* gene is highly polymorphic. For this gene, fifty SNPs and three insertions/deletions have been reported. The ABCB1 protein is encoded by the *ABCB1* gene.[62] Overexpression of the *ABCB1* gene in cancer cells induces resistance to chemotherapeutic agents. Distribution of some allelic variants appears to be co-dependent on similar transporter genetic variations.[31,32] SLCO1B3 is a solute carrier organic protein encoded by the *SLCO1B3* gene. The solute vector superfamily is responsible for the uptake and transfer of drugs and is classified into 46 subfamilies, which includes the organic anion transporting polypeptides (OATPs), organic anion transporters (OAT) and organic cation transporter (OCT) subfamilies.[31,32] OATPs are a large family of membrane-bound influx transporters that are responsible for the cellular uptake of a wide range of endogenous substances such as bile salts and hormones, as well as exogenous substances including bile salts, hormones and clinically administered drugs such as antibiotics, cardiac glycosides and anticancer agents, which are administered clinically.[31,32]

2.7.5 Effect of polymorphisms on DNA repair pathways

There are four major DNA repair pathways: nucleotide excision repair (NER), base excision repair (BER), double-strand break repair (DSB) and mismatch repair (MMR).[64] The NER pathway is particularly important in excision mechanisms involved in the removal of damaged DNA. Damage to DNA results in DNA adducts occurring in the form of thymine dimers and 6,4-photoproducts.[64] Recognition of the DNA damage leads to the removal of short single-stranded DNA segments containing the lesion. The latter has been found applicable to platinum chemotherapy drugs.[64,65] Three genes in the NER pathway have been identified as role players in the pharmacogenetics of platinum compounds; excision repair cross-complementing 1 (*ERCC1*), xeroderma pigmentosum group D (*XPD*) and xeroderma pigmentosum group G (*XPG*).[64]

2.8 Drug label regulations

Advances in next-generation sequencing techniques have accelerated pharmacogenomics knowledge. To date over 200 genome-wide association studies (GWAS) relating to pharmacotherapy response have been published.[66] The effects of genetic variation have not gone unnoticed by the US Food and Drug Administration (FDA) and European Medicines Agency (EMA). In 2005 the FDA released a framework for developing and approving pharmacogenetics information and in 2007, the first drug label was revised to include pharmacogenetic information, the label was for the drug warfarin, and included the potential impact of *CYP2C9 and VKORC1* genetic variations on dosing requirements and risks.[67] Variability in drug efficacy and adverse side effects can be explained by genetic variation in approximately 80% of cases. Over 400 genes are consider clinically relevant to drug metabolism and approximately 200 pharma genes are associated with ADRs.[68] Prior knowledge of these effects would be beneficial when selecting and determining dosing requirements of medication. Cancer drugs maintain a strong presence in pharmacogenetics and targeted therapies. Improved knowledge of gene-drug interactions allows for the prediction of treatment efficacy in targeted patient subgroups.[66] Although pharmacogenetic data is available, the clinical implementation thereof is underutilized. Furthermore, most clinical trials are performed solely on the intended target market,

even when the drug is intended to be used for individuals of diverse ethnicities. The effect thereof is an increasing risk for unforeseen drug-related toxicities in other populations. Many drug and clinical trials are thus not conducted in under studied populations such as in sub-Saharan Africa. An example of such a drug that showed varied response is that of 5-Fluorouracil, the side effect of which was hematologic toxicities, which were more prevalent in populations of African descent.[69] There is a desperate need to bridge this gap, and to better understand how the drug will affect other subgroups.

2.9 Consortia on pharmacogenomics research

Implementing pharmacogenetic knowledge in clinical practice requires evidence based guidelines. Many committees have been established, the most commonly referenced is the Dutch Pharmacogenetics Working Group (DPWG) and the Clinical Pharmacogenetics Implementation Consortium (CPIC,) both with a main focus on pharmacotherapeutic recommendation in combination with a patient's genotype or predicted phenotype.[70] The DPWG has a five-point-scale scoring system used for the level of evidence (0-4) and a seven-point scale (AA# - F) for clinical relevance or impact. The overall score of each combination is the highest level of evidence and the highest level of relevance assigned. The CPIC's grading system consists of three levels, high (consistent results from well-designed, well-conducted studies), moderate (evidence is sufficient to determine effects, but strength of the evidence in limited by the number, quality or consistency of the study), or weak (evidence is insufficient to assess the effects of health outcomes).[70]

2.10 Cancer treatment and gene-drug interactions

2.10.1 Androgen-deprivation therapy

The hormone, testosterone, serves as the main fuel source for PrCa cell growth and is a common target for therapeutic intervention.[23] ADT inhibits the release of testosterone, thereby starving the prostate cell of this fuel source.[23]

Abiraterone acetate, which is sold under the trade name Zytiga®, is an anti-androgen medication and a derivative of steroidal progesterone.[71] It is administered orally as a salt prodrug in this form as, it is less susceptible to hydrolysis by esterases, thereby

increasing the bioavailability of the drug. Abiraterone acts by inhibiting the steroidal enzyme CYP17A1 (17 alpha-hydroxylase/C17,20 lyase). *CYP17A1* encodes for the enzyme that catalyzes the biosynthesis of androgen by inhibiting the conversion of 17-hydroxyprognenolone to dehydroepiandrosterone (DHEA), lowering the levels of testosterone and other androgens in the serum. Abiraterone is metabolised into inactive metabolites by *CYP3A4* and *SULT2A1*. Blockage of 17α-hydroxylase activity, results in the accumulation of upstream mineralocorticoids such as 11-deoxycorticosterone, leading to secondary hyperaldosteronism, which includes fluid retention and hypokalemia.[72] *HSD3B1* encoding 3β-hydroxysteroid dehydrogenase-1 is expressed mainly in peripheral tissues (including prostate and breast) and is a rate-limiting enzyme required for all pathways of dihydrotestosterone synthesis.[72]

Polymorphisms render *HSB3B1* resistant to proteasomal degradation resulting in the accumulation of this enzyme and gain of function.[72] Other genes such as the testis-specific Y-encoded-like protein (*TSPYL)* have been reported to regulate expression of many CYP genes including *CYP3A* and *CYP17A1*, both key enzymes in androgen biosynthesis.[72] *TSPYL1* rs3828743 SNP has been shown to diminish TSPYL1's ability to supress *CYP3A4* expression, thereby decreasing abiraterone's concentration and increasing cell proliferation and thus worsening the repose to abiraterone acetate as well as decreasing progression-free survival (Figure 6).[73]

2.10.2 Selective oestrogen receptor modulators

More than 50% of primary BrCas express hormone receptors (oestrogen receptor (ER) and/or progesterone receptor (PR)), and therefore endocrine therapy like TAM and aromatase inhibitors are thus offered for treatment.[74] Women with early stage BrCa are likely to receive adjuvant TAM for 5 years, while postmenopausal women are likely to be offered aromatase inhibitors.[74]
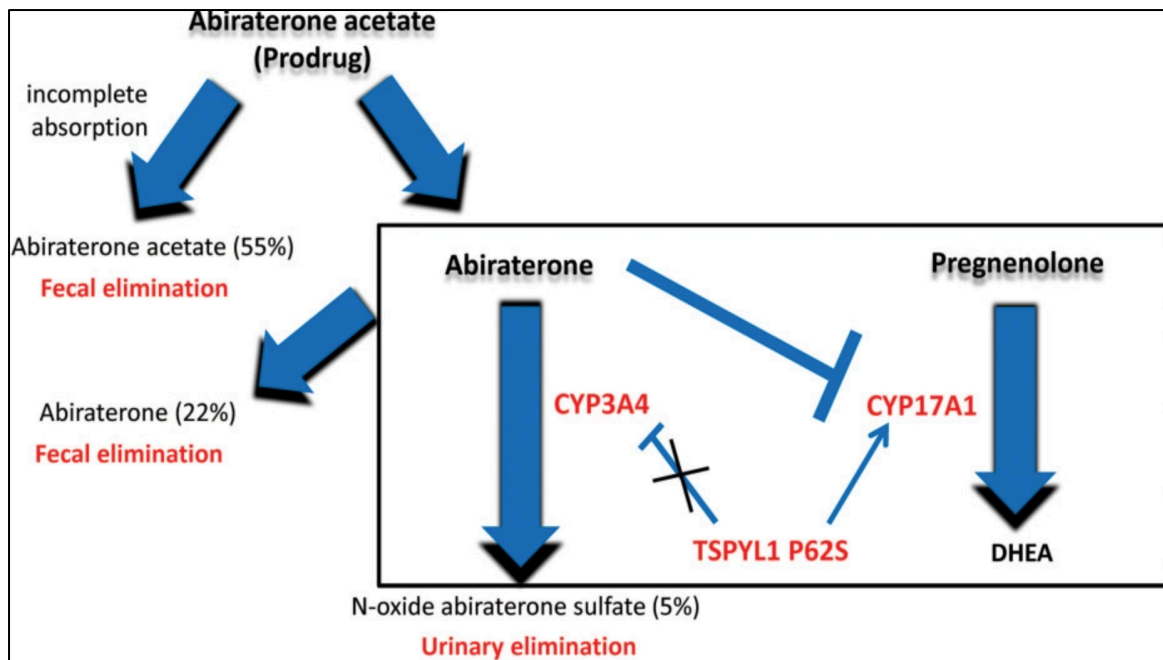
Figure 6. Schematic of the postulated schematic effect of TSPYL1 SNP affecting the efficacy of abiraterone acetate. *TSPYL1* rs3828732 SNP affects the suppression of *CYP3A4*, causing an increase in the expression of CYP3A4 and lowering abiraterone exposure, which affects the inhibition of abiraterone on CYP17A1 and further drug response (Reproduced with permission[73]).

TAM is a nonsteroidal agent that binds to ERs, inducing a conformational change in the receptor, thereby modulating oestrogen-induced transcription.[75] This drug has tissue-specific activity and is used in the treatment and prevention of oestrogen-dependent BrCa.[76] TAM has both estrogenic and anti-oestrogenic effects. It acts as an anti-oestrogen inhibiting agent in the mammary tissue and as an oestrogen stimulating agent in cholesterol metabolism. This results in a blockage or change in the expression of oestrogen dependent genes.[45] The prolonged binding of TAM to the nuclear chromatin results in reduced DNA polymerase activity, impaired thymidine utilization, blockade of oestradiol uptake and decreased oestrogen response.[45] It is likely that TAM interacts with other co-activators or corepressors in the tissue and binds to different ERs (ER-alpha or ER-beta), producing both oestrogenic and anti-oestrogenic effects.[45]

TAM is a prodrug and is extensively metabolized after oral administration into two metabolites in plasma; N-desmethyl-4-hydroxy-tamoxifen (endoxifen), the major metabolite and 4-hydroxy-tamoxifen (4-OH-TAM) the secondary metabolite (Figure 7). 4-OH-TAM formation is catalysed mainly by CYP2D6 and also by CYP2C9 and CYP3A4 but to a lesser extent.[76] At high TAM concentrations, CYP2B6 also catalyses 4-hydroxylation of the parent drug.[76] 4-OH-TAM possesses 30- to 100-fold greater affinity for the ER and 30- to 100-fold greater potency for inhibiting oestrogen-dependent cell proliferation than TAM.[76] Symptoms associated with TAM toxicity includes respiratory difficulties and convulsions.[45,77] 4-OH-TAM and endoxifen are further glucuronidated and sulphated in phase 2 reactions and catalyzed by urifine-5-diphosphoglucuronosyltransferases (UGTs) and sulfonyl-transferases (SULTs).[77]
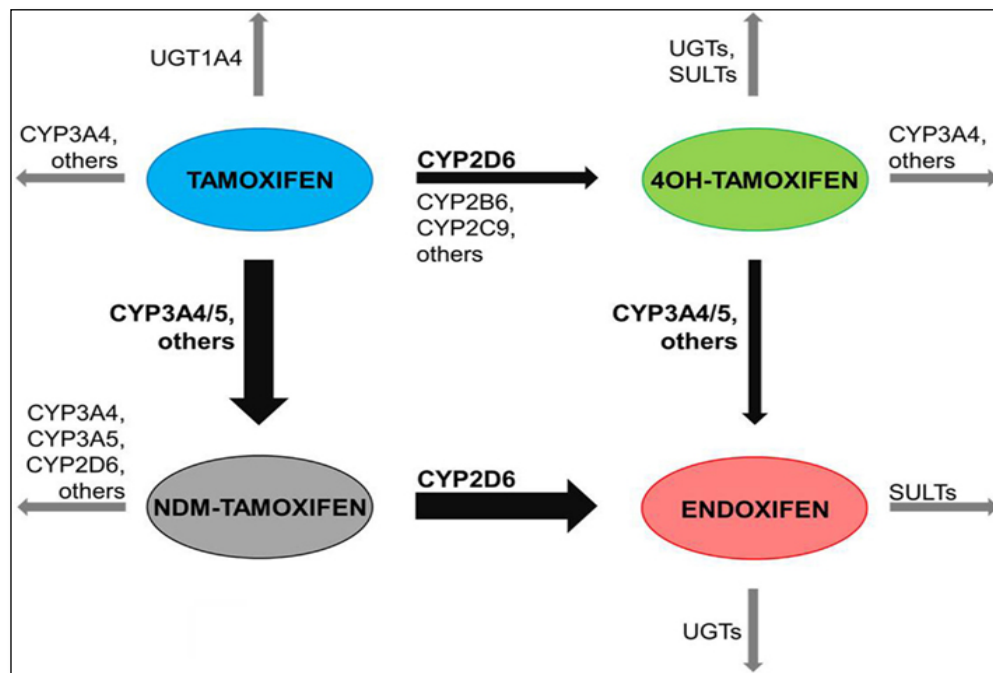
Figure 7. Main biotransformation route of tamoxifen (Reproduced with permission[77]). CYPs: Cytochrome P450 subfamilies, UGTs: diphosphoglucuronosyltransferases, SULTs: sulfonyl-transferases.

Aromatase inhibitors block the conversion of androgens to oestrogens, depleting the circulation of the latter.[45] Letrozole (trade name Femara®) is an oral non-steroidal aromatase inhibitor which is used as an adjuvant treatment for hormonally-responsive BrCa and a first-line treatment for postmenopausal women with hormone receptor positive or hormone receptor unknown locally advanced or metastatic BrCa.[61] It is also

indicated for the treatment of advanced BrCa in postmenopausal women with disease progression following antioestrogen therapy. The treatment response rate to aromatase inhibitors varies between 35 and 70%.[78]

Oestrogens are produced by the conversion of androgens via the activity of the aromatase enzyme.[61] Several enzymes are involved in the formation and metabolism of oestrogen, and also the metabolism of aromatase inhibitors, namely *CYP19A1, CYP1A1, CYP1A2, CYP2C9, COMT* and *CYP3A4.*[61] Aromatase is encoded by the *CYP19A1* gene located on chromosome 15q21.1.[60] Higher levels of aromatase have been found in breast tumours than normal breast tissue.[79] Letrozoleis highly specific for the aromatase enzyme and blocks the production of oestrogens by competitive, reversible binding to the heme of its CYP unit, resulting in a complete depletion of oestrogen.[36,79] Breast tissue is stimulated by oestrogen, and therefore a decrease in the production of oestrogen results in a decrease in the recurrence of breast tumour tissue.[80]

As letrozole is not a steroid, it does not affect the production of mineralocorticoids or corticosteroids. Side effects include signs and symptoms of hypoestrogenism.[81] There is the concern that long-term use may lead to osteoporosis and it is often prescribed with osteoporosis-fighting medication and vitamin D.[81] Letrozole has been shown to reduce oestrogen levels by 98%, while resulting in a two-fold increase in testosterone levels.[81]

Genetic polymorphisms in the *CYP19A1* gene have been shown to alter sex hormone levels, which could potentially explain the increased risk of BrCa in individuals exposed to oestrogen (Figure 8).[82] It is reported that the rs4646 polymorphism can be used to identify the subgroup of patients likely to respond poorly to neoadjuvant letrozole treatment, thereby posing a poor prognosis.[60] Letrozole is primarily metabolised by the hepatic system via *CYP3A4* and *CYP2A6,* to an inactive metabolite whose glucuronide conjugate is mainly excreted via the kidney.[83]
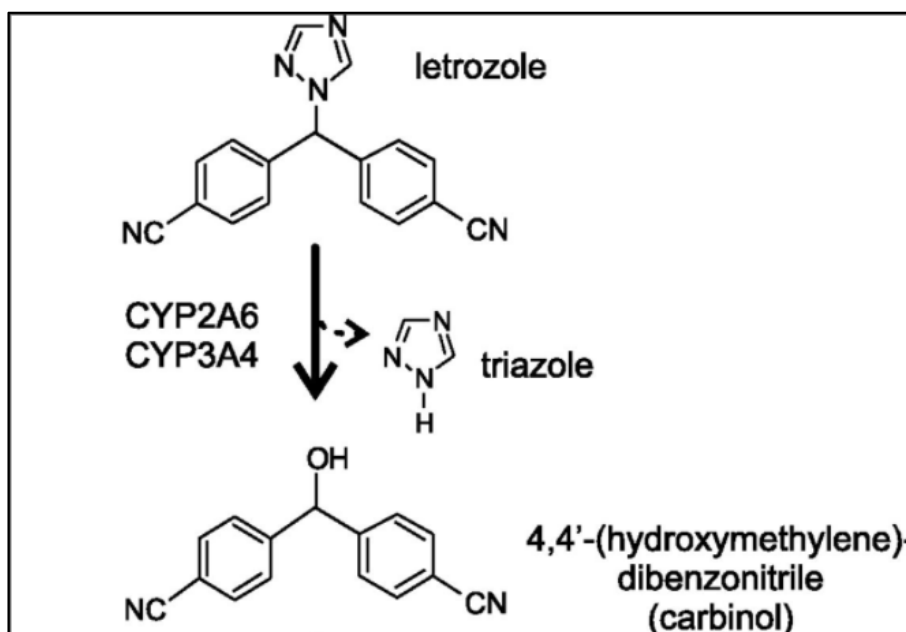
Figure 8. The metabolism of letrozole to carbinol by CYP2A6 and CYP3A4 (Reproduced with permission[83]).

2.10.3 Chemotherapy and other drugs for adjuvant therapy

2.10.3.1 Alkylating agents

Alkylating agents work via three different mechanisms to disrupt DNA function and cause cell death; i) attachment of alkyl groups to DNA bases, resulting is the DNA being fragmented by repair enzymes in their attempts to replace the alkylated bases, thus preventing DNA synthesis and RNA transcription from the affected DNA, ii) by DNA damage via the formation of intra-strand and inter-strand cross-links (bonds between atoms in the DNA) which prevents DNA from being separated, blocking cellular processes such as replication and transcription, and iii) by induction of mispairing of the nucleotides, leading to mutations.[84]

Cyclophosphamide is an antineoplastic in the class of alkylating agents and is used to treat various forms of cancer, including carcinoma of the breast. Metabolism and activation occur in the liver.[85] The *CYP2B6* isoform is the enzyme with the highest 4-hydroxylase activity.[85] Cyclophosphamide undergoes activation to form the metabolites, phosphoramide mustard and acrolein.[85] Cyclophosphamide appears to induce its own metabolism which results in an overall increase in clearance, increase formation of 4-hydroxyl metabolites and shortened half-life ($t_{1/2}$) values following

repeated administration.[85] Adverse reactions include neutropenia, febrile neutropenia, fever, alopecia, nausea, vomiting and diarrhoea.[85] Multiple CYP enzymes have been implicated in the metabolic activation of cyclophosphamide, namely; *CYP2A6, CYP2B6, CYP2C9* and *CYP3A4/5*, (Figure 9).[85]
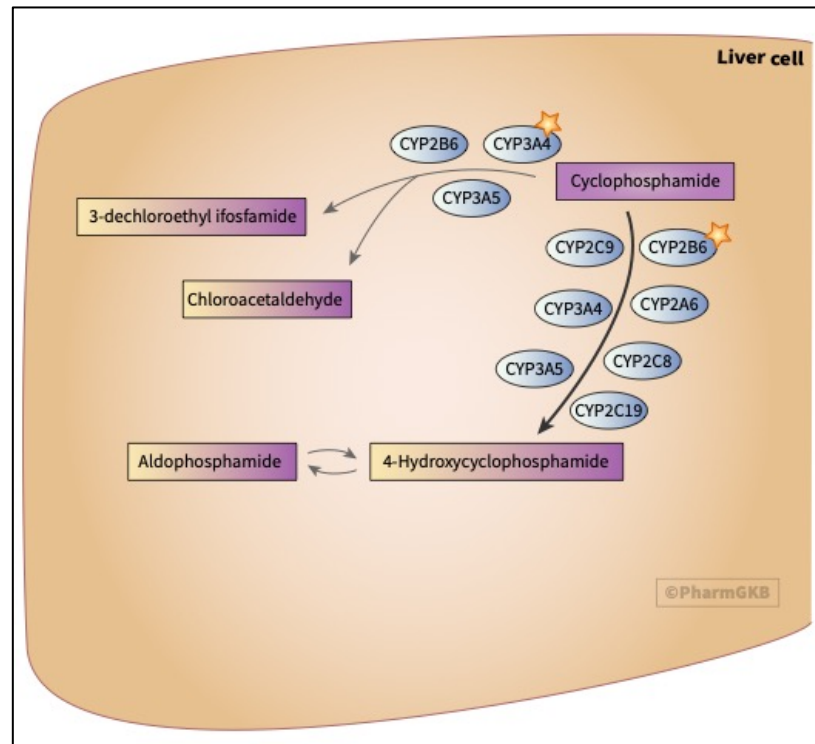
Figure 9. The genes involved in the metabolism pathway of cyclophosphamide in human liver cells (Reproduced with permission[86]).

Another group of alkylating agents are the platinum compounds; cisplatin, carboplatin and oxaliplatin, which are used to treat metastatic testicular tumours, metastatic ovarian tumours and advanced cancer of the bladder.[34] The effectiveness of platinum agents is compromised as a result of decreased accumulation of the drug, detoxification, no DNA adduct formation and increased DNA repair activity.[34] Although effective, these drugs result in severe side effects such as nephrotoxicity, hematogenesis, ototoxicity and neurotoxicity.[84]

Genetic polymorphisms in DNA repair enzymes and detoxification have an effect on the efficacy and toxicity of these drugs.[65,84] Polymorphisms in glutathione dependent enzymes can influence treatment response to platinum compounds.[34] Improved DNA repair machinery caused by polymorphisms in X-ray repair cross-complementing

protein 1 (*XRCC1*) and excision repair cross-complementing 1 (*ERCC1*) genes are known to affect the drug's effectiveness.[34] *ERCC1* is involved in excision repair on the nucleus of the cell (Figure 10).[87]
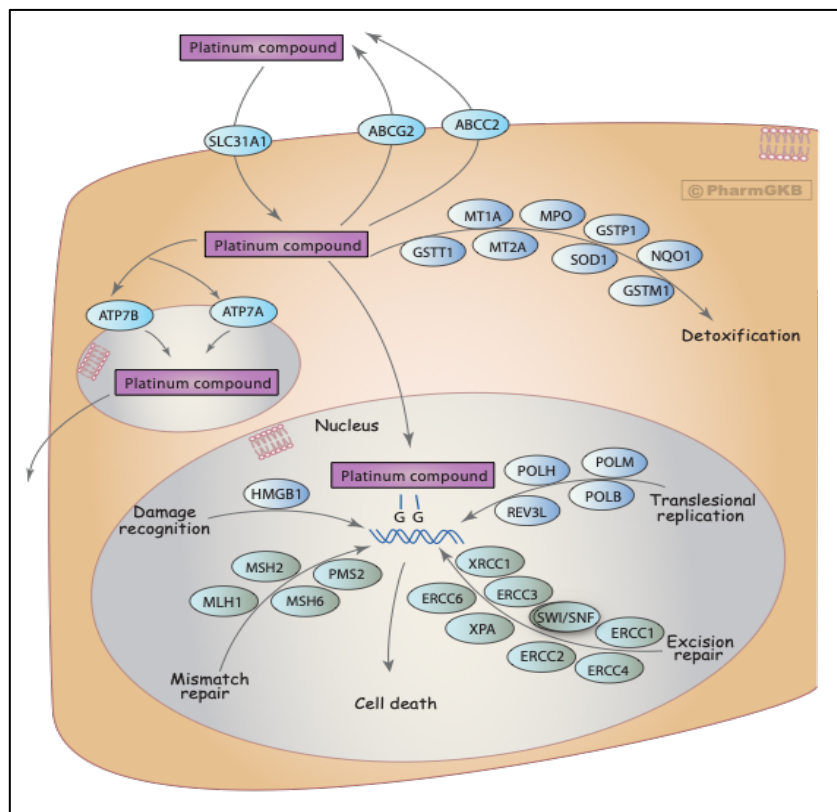
Figure 10. Candidate genes involved in the metabolism of platinum drugs (Reproduced with permission[87]).

2.10.3.2 Anthracyclines

General properties of anthracyclines include interaction with DNA as well as intercalation (squeezing between the base pairs), DNA strand breakage and inhibition of the enzyme topoisomerase II.[45] Most of these compounds originate from natural sources, however, they lack the specificity of the antimicrobial antibiotics and thus produce significant toxicity. The anthracyclines are among the most important antitumor drugs currently available on the market.[45]

Doxorubicin is an example of a cytotoxic anthracycline antibiotic isolated from cultures of *Streptomyces peucetius var. caesius*.[85] It binds to nucleic acids, presumably by specific intercalation of the planar anthracycline nucleus with the DNA double helix.[85]

It is reported to inhibit polymerase activity, affect regulation of gene expression and produce free radical damage to DNA.[85] It is used to produce regression in disseminated neoplastic conditions, among other BrCa. Doxorubicin is also indicated for use as a component of adjuvant therapy in women with evidence of axillary lymph node involvement following resection of primary BrCa.[85]

Doxorubicin is capable of undergoing three metabolic fates; one-electron reduction, two-electron reduction and deglycosidation.[85] However, approximately half of the dose is eliminated from the body unchanged. A two electron reduction yields doxorubicinol, a secondary alcohol, which is considered the primary metabolic pathway. The one electron reduction is facilitated by several oxidoreductases to form a doxirubicin-semiquinone radical. These enzymes include mitochondrial and cystolic NADPH dehydrogenates, xanthine oxidase and nitric oxide synthases.[85] Deglycosidation is a minor metabolic pathway (1-2% of the dose undergoes this pathway) with the resultant metabolites deoxyaglycone or hydroxyaglycone formed via reduction or hydrolysis, respectively. Enzymes involved with this pathway include xanthine oxidase, NADPH-cytochrome P450 reductase and cytosolic NADPH dehydrogenase.[85] Glutathione-S-transferase and multi drug resistance (MDR) genes may contribute to doxorubicin resistance.[85] Several transporters are responsible for the transportation of doxorubicin and include *ABCB1, ABCC1, ABCC2, ABCG2, RALBP1* and *SCL22A16,* (Figure 11). Genetic variations in transporter genes have been associated with drug resistance.[88]

Epirubicin, the 4'-epi-isomer of doxorubicin, is another example of an anthracycline. This compound exerts its antitumor effects by interfering with the synthesis and functioning of DNA.[85] It is used as a component of adjuvant therapy in patients with evidence of axillary node tumour involvement following resection of primary BrCa.[85] Although epirubicin is extensively and rapidly metabolized in the liver, it is also metabolized by other organs and cells, including red blood cells.[85]
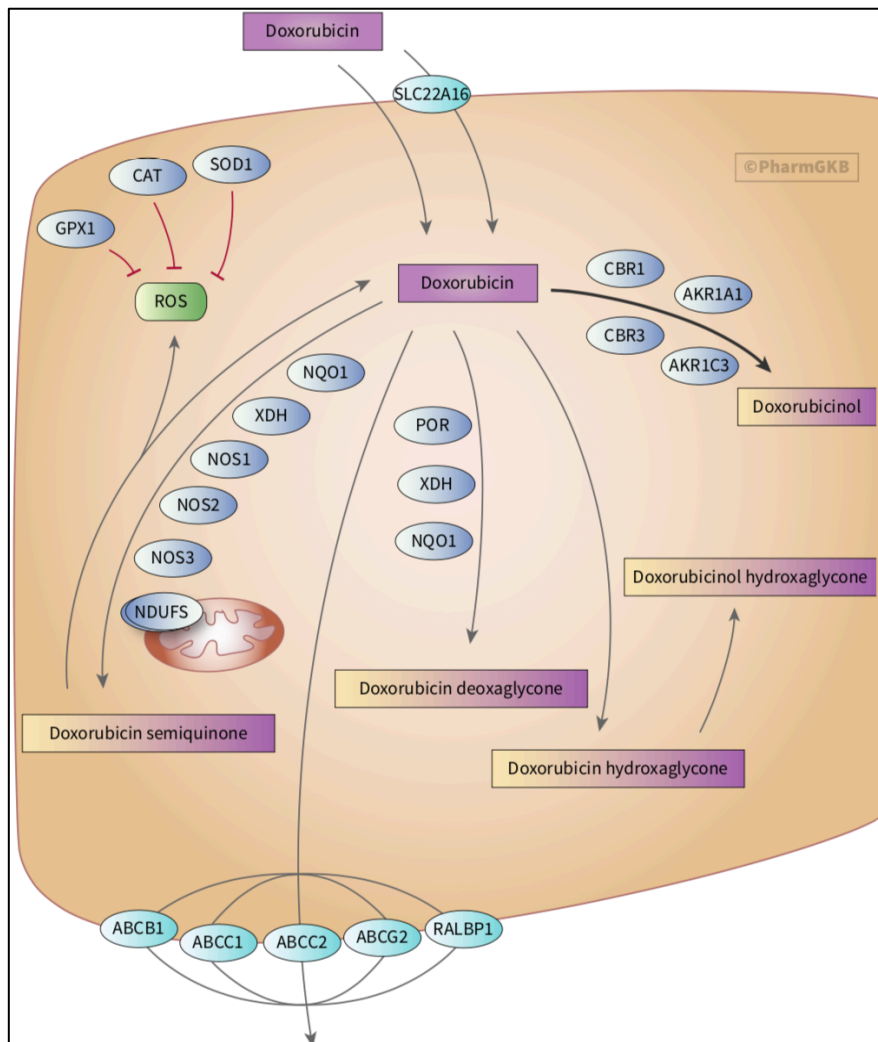
Figure 11. Transport and metabolism of doxorubicin reflecting candidate genes in the metabolic pathway (Reproduced with permission[88]).

The four main metabolic routes are: i) reduction of the C-13 keto-group with the formation of the 13(S)-dihydro derivative, epirubicinol; ii) conjugation of both the unchanged drug and epirubicinol with glucuronic acid; iii) loss of the amino sugar moiety through a hydrolytic process with the formation of the doxorubicin and doxorubicinol aglycones; and iv) loss of the amino sugar moiety through a redox process with the formation of the 7-deoxy-doxorubicin aglycone and 7-deoxy-doxorubicinol aglycone.[85] Although epirubicinol exhibits *in vitro* cytoxic activity (~10% that of epirubicin), it is unlikely to reach sufficient concentrations *in vivo* to produce similar effects. Bone marrow aplasia, grade 4 mucositis and gastrointestinal bleeding are reported side effects of the drug. Formation of epirubicin glucuronide by liver UDP-glucuronosyltransferase (UGT) is the primary inactivating pathway.[85]

2.10.3.3 Taxane derivatives

Taxane derivatives interfere with microtubule growth, by arresting their function by means of hyper-stabilizing their structure. This destroys the cell's ability to use its cytoskeleton in a flexible manner.[89] Taxanes bind to the β-subunit of tubulin, the "building block" of microtubules, locking these building blocks in place. The resulting microtubule/taxane complex does therefore not have the ability to disassemble.[89] This adversely affects cell function, since shortening and lengthening of microtubules (termed dynamic instability) is necessary for their function as a transportation highway for the cell.[89] Taxanes have a narrow therapeutic index with several adverse side effects, including hematopoietic and neurologic toxicities.[85,89]

Docetaxel is an anti-mitotic chemotherapy medication used mainly for the treatment of breast (locally advanced or metastatic BrCa after failure of prior chemotherapy), ovarian and non-small cell lung cancer.[62,89] It is also used in combination with prednisone, in the treatment of patients with androgen independent (hormone refractory) metastatic PrCa.[62,89] *In vitro* drug interaction studies revealed that docetaxel is metabolized by the CYP3A4 isoenzyme (Figure 12). Mutations in transport proteins, such as ATP-binding cassettes (ABC) and solute vector carriers (SLC) have been shown to affect the intracellular concentration of the drugs. Over dosing with this drug results in bone marrow suppression, peripheral neurotoxicity and mucositis.[62,89]

Paclitaxel is a chemotherapeutic agent that was first isolated in 1971 from the bark of the Pacific yew tree. It is available as an intravenous solution for injection and the newer formulation contains albumin-bound paclitaxel marketed under the brand name Abraxane®.[89] Abraxane® is specifically indicated for the treatment of metastatic BrCa and locally advanced or metastatic non-small cell lung cancer. It is a novel anti-microtubule agent that promotes the assembly of microtubules from tubulin dimers and stabilizes microtubules by preventing depolymerization.[89] This results in the inhibition of the normal dynamic reorganization of the microtubule network that is essential for vital interphase and mitotic cellular functions. In addition, paclitaxel induces abnormal arrays or "bundles" of microtubules throughout the cell cycle and multiple asters of

microtubules during mitosis. Paclitaxel is metabolized primarily to 6-alpha-hydroxypaclitaxel by the isoenzyme CYP2C8; and two minor metabolites, 3'-*p*-hydroxypaclitaxel and 6-alpha, 3'-*p*-dihydroxypaclitaxel, by CYP3A4 in the liver (Figure 12). Symptoms of overdose are similar to docetaxel.[89]
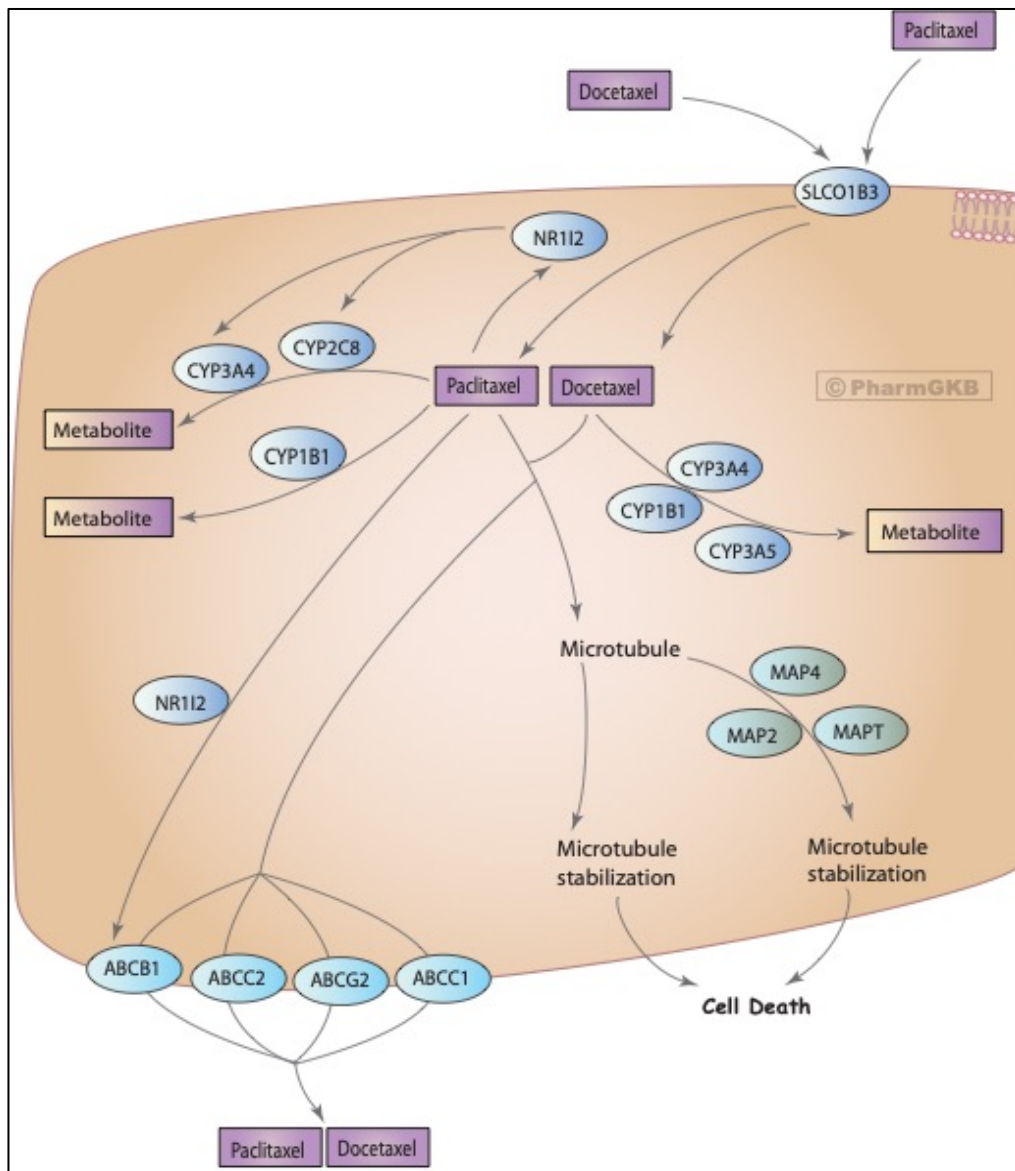
Figure 12. Genes involved in the metabolism and transport of paclitaxel and docetaxel, along with the downstream effect of the drugs (Reproduced with permission[32]).

Cabazitaxel is another anti-neoplastic used in conjunction with the steroid, prednisone. Cabazitaxel is used to treat hormone refractory metastatic PrCa.[89] This drug is prepared by semi-synthesis with a precursor extracted from yew needles (10-

deacetylbaccatin III). Cabazitaxel binds to tubulin and promotes its assembly into microtubules while simultaneously inhibiting disassembly.[89] This leads to the stabilization of microtubules, which results in the interference of mitotic and interphase cellular functions. The cell is then unable to progress further into the cell cycle, being stalled at metaphase, thus triggering apoptosis of the cancer cell. Cabazitaxel is extensively metabolized in the liver (>95%), mainly by the CYP3A4/5 isoenzyme (80% to 90%) and to a lesser extent by CYP2C8, which results in 20 metabolites.[89] Cabazitaxel may cause serious side effects including neutropenia, hypersensitivity reactions, gastrointestinal symptoms and renal failure. Overdosing with this drug results in bone marrow suppression and gastrointestinal disorders.[89]

2.10.3.4 <u>Pyrimidine antimetabolites</u>

Antimetabolites masquerade as purine or pyrimidine, the building blocks of DNA.[85] Thereby preventing their incorporation into DNA during the "S" phase (of the cell cycle), and halting normal cell development and division.[85]

Fluorouracil, also known as 5-FU, is one example of an antimetabolite. Fluorouracil blocks the enzyme which converts the cytosine nucleotide into the deoxy derivative. In addition, DNA synthesis is further inhibited since fluorouracil blocks the incorporation of the thymidine nucleotide into the DNA strand.[90] The precise mechanism of action of fluorouracil has not been fully determined, but it is thought to be via the binding of the deoxyribonucleotide of the drug (FdUMP) and the folate cofactor, N5–10-methylenetetrahydrofolate, to thymidylate synthase (TS) to form a covalently bound ternary complex. This results in the inhibition of the formation of thymidylate from uracil, which leads to the inhibition of DNA and RNA synthesis and cell death. Fluorouracil can also be incorporated into RNA in place of uridine triphosphate (UTP), producing a fraudulent RNA and interfering with RNA processing and protein synthesis.[90]

The metabolism of fluorouracil takes place in the liver, resulting in inactive degradation products (e.g. CO_2, urea and α-fluoro-ß-alanine).[85] The dihydropyrimidine dehydrogenase (DPD) enzyme is responsible for the formation of the inactive

metabolite dehydro-fluorouracil from 5-FU (Figure 13). DPD metabolises 80-85% of 5-FU. DPD activity is diminished in 3-5% of individuals.[85] To date more than 30 genetic variations in *DPYD* have been reported, all of which result in decreased enzyme activity. Mutations in this gene have been associated with drug toxicity.[91] Thymidylate synthase (TS) a folate dependent enzyme which plays a role in cellular expression of several genes and affects cell proliferation and death, is the target of 5-FU. Changes in gene expression and function may result in altered chemosensitivity.[85]

Capecitabine is a newer orally administered formulation of 5-FU used in the treatment of metastatic BrCa which is resistant to both paclitaxel and an anthracycline-containing chemotherapy regimen.[85] This drug may also be used in combination with docetaxel for the treatment of metastatic BrCa in patients who have failed to respond to or recurred or relapsed during or following anthracycline-containing chemotherapy. Capecitabine is a prodrug that is enzymatically converted to fluorouracil (antimetabolite) in the tumour by thymidine phosphorylase, where it inhibits DNA synthesis and slows growth of tumour tissue.[85] Fluorouracil is further metabolized in both normal and tumour cells to two active metabolites 5-fluoro-2′-deoxyuridine 5′-monophosphate (FdUMP) and 5-fluorouridine triphosphate (FUTP). These metabolites cause cell injury by two different mechanisms. One where, FdUMP and the folate cofactor, N5-10-methylenetetrahydrofolate, bind to TS to form a covalently bound ternary complex.[85] This binding inhibits the formation of thymidylate from 2'-deoxyuridylate. Thymidylate is the necessary precursor of thymidine triphosphate, which is essential for the synthesis of DNA, therefore a deficiency of this compound can inhibit cell division. Alternatively, nuclear transcriptional enzymes can mistakenly incorporate FUTP instead of uridine triphosphate (UTP) during the synthesis of RNA. This metabolic error can interfere with RNA processing and protein synthesis through the production of fraudulent RNA.[85]

Figure 13. Metabolic pathway of fluoropyrimindines (Reproduced with permission[92]).

2.10.3.5 Folate antimetabolites

These drug classes are antineoplastic, antimetabolites with immunosuppressant properties. They are inhibitors of tetrahydrofolate dehydrogenase and prevent the formation of tetrahydrofolate, necessary for synthesis of thymidylate, an essential component of DNA.[85] An example of a folate antimetabolite is methotrexate (MXT).[85] MTX is used alone or in combination with other anticancer agents in the treatment of BrCa, epidermoid cancers of the head and neck, advanced mycosis fungoides (cutaneous T cell lymphoma) and lung cancer, particularly squamous cell and small cell types. The mechanism of action is similar to what has been described above for

5-FU. MTX inhibits folic acid reductase which is responsible for the conversion of folic acid to tetrahydrofolic acid, thereby inhibiting DNA synthesis and cellular replication.[85] At two stages in the biosynthesis of purines and at one stage in the synthesis of pyrimidines, one-carbon transfer reactions occur which require synthesis of specific coenzymes in the cell from tetrahydrofolic acid.[93] Tetrahydrofolic acid itself is synthesized in the cell from folic acid with the help of the enzyme, folic acid reductase. MTX mimics folic acid, therefore binding to it relatively strongly, and inhibiting enzyme binding.[93] Thus, DNA synthesis cannot proceed because the coenzymes needed for the one-carbon transfer reactions are not produced from tetrahydrofolic acid, due to its absence. MTX selectively affects the most rapidly dividing cells (neoplastic and psoriatic cells), by inhibiting folate targets. The bioavailability of MTX is affected by several ABC transporters (*ABCC2, ABCB1, ABCG2*) which exert an influence on either the movement of MTX out of the cells and back into the intestinal tract or into the blood (*ABC1* and *ABCC3*) (Figure 14).[93] Systemic clearance of MTX happens primarily via renal glomerular filtration. Several rental transporter proteins have an affinity for MTX (*SLC226, SLC22A8, SLC19A1, ABCG2, ABCC2* and *ABCC4*). Also, SNPs in *ABCC2* have been associated with delaying MTX clearance.[93]

MTX undergoes hepatic and intracellular metabolism, resulting in polyglutamated forms which can be converted back to MTX by hydroxylase enzymes.[85] MTX clearance rates vary widely and are generally decreased at higher doses. Delayed drug clearance has been identified as one of the major factors responsible for MTX toxicity.[85] Methylenetetrahydrofolate reductase (MTHFR) regulates the pool of intracellular folate, and polymorphisms in the *MTHFR* gene may result in differential toxicity.[85]

2.11 Summary of clinical pharmacogenetic studies in African populations

Most patient-based studies and drug development programs have been conducted on populations of European and Asian ancestry. The African continent, although considered the cradle of human origin and presenting with high genetic diversity and a complex population structure, has to a large extent been under studied. The genetic diversity within the African continent has been highlighted in several studies.
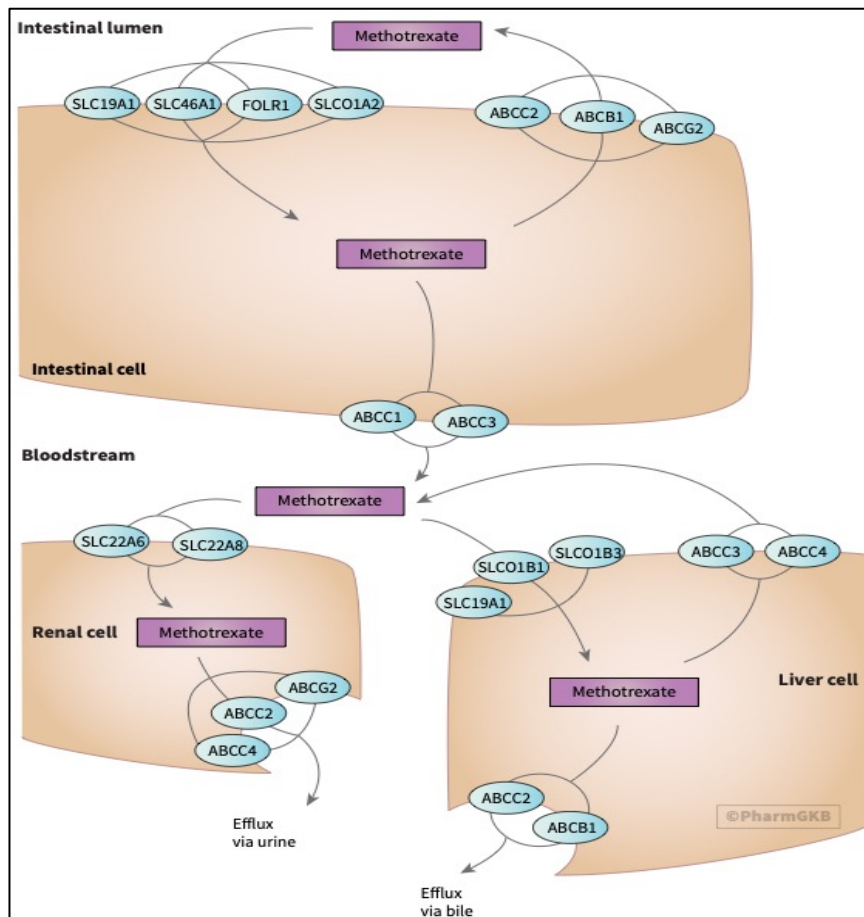
Figure 14. Schematic representation indicating the uptake, transport and elimination of methotrexate (Reproduced with permission[93]).

Given the high burden of disease in the region, HIV, TB and cancer, optimal drug therapies are essential to lower treatment costs and hospital admission due to ADRs. Most pharmacogenetic studies that have been conducted in African populations relate to genotype-phenotype associations and drug response, and have focused on North Africa.[94] Cardiovascular disease, kidney disease, infectious diseases and mental disorders are the among the most frequently studied.[94] Below is a summary of studies that have included the sub-Saharan African population.

A study by da Rocha *at al.* looked at the ADRs caused by fluoropyrimidine in a sub-Saharan African Population.[95] It was noted that the *DPYD* gene variation (rs115232898 – C), was linked to severe toxicity.[95] A second variant which has been less studied, rs2297595 – C, showed a significant allele frequency difference between

the sub-Saharan African population and the East African population. The Southern African population was represented by 100 samples, 12.1% of the total study cohort.[95]

Rajman *et al.* in two separate articles highlighted the genetic diversity in Africa, compared to Asian and Caucasian populations.[59] This genetic diversity was first highlighted in cytochrome P450 genetic variants, and then in the drug transporter genes, adenosine triphosphate-binding cassette (ABC) and solute carrier (SLC) superfamilies.[96] The pharmacological implications of genetic diversity in the populations of sub-Saharan African ancestry was highlight by Tshabalala *et al.* through ultra-deep sequencing of a South African Bantu-speaking cohort (n = 40).[9] Novel variants were noted in 65 pharmacologically-related genes.[9] Ultra-deep sequencing allowed for the identification of the novel gene variants, which would likely not have been detected through other genotyping techniques.

Another study highlighted the diversity of variants important in predicting pharmacogenetic-based warfarin dosing in African populations, including variants reported in CYP2C9 and VKRO1.[97] The study called for more African population based studies to enable the identification of all possible biomarkers for drug response prediction.[97] The extent of genetic diversity within ADME genes within the Southern African populations is evident. The African ADME landscape is unique, and requires further characterization in order to develop precision medicine guidelines and tools specific to the African population. Larger population based studies are required, with the inclusion of larger Southern African cohorts. The intra-population variation was also highlighted in a study by da Rocha *et al.*, using data from 485 high-coverage whole genome sequencing data, of which 210 samples were from South Africa, Namibia and Botswana.[98] The limitation of array-based genotyping technologies applied as high throughput methods was highlighted, and it was proposed that targeted sequencing or whole-exome sequencing might be better suited for characterising ADME genes and identifying rare variants.[98] However, the cost of sequencing remains too high to conduct large cohort sequencing projects.

A clinical pharmacogenetic study aimed to determine the concentration time curve and neutrophil toxicity in Ethiopian breast cancer patients receiving cyclophosphoamide. *CYP3A5*3/*6* carriers showed a decreased clearance of cyclophosphoamide, while

CYP2C9*2/*3 carriers indicated an increased clearance. Inter-individual variation was also reported as a result of other factors such as body surface area, body-mass index and dosage regime.[99]

To date, there are still only a limited number of published articles relating to the sub-Saharan African populations. Many of the studies mentioned above have limited sample numbers. Larger cohort studies are required, with a particular focus on the sub-Saharan African population, where non-optimal treatment outcomes amplify the burden of disease.

2.12   Study aim and objectives

The aim of the study was to determine the prevalence of genetic polymorphisms important in the safety and efficacious treatment outcomes of anti-cancer drugs in the South African Bantu speaking population.

There were four primary objectives:

1. To determine genotypic and allele frequencies of sixty pre-selected variants in a diverse South African Bantu speaking populations.
   a. The secondary objectives were:
      i. To genotype a large study cohort using a high throughput multiplexed array.
      ii. To validate the accuracy of the TaqMan® multiplexed OpenArray™.
      iii. To ensure the absence of data set error, by performing a Hardy Weinberg equilibrium test.
2. To compare intra population frequencies of the selected pharmacogenetic variants using the 1000 Genomes database for African and European ancestry.
3. To compare inter-population frequencies of the selected pharmacogenetic variants from the data generated.
4. To predict the clinical implication of the selected pharmacogenes in the South African Bantu speaking populations based on the prevalence of the risk allele withing the study population.

CHAPTER 3: MATERIALS AND METHODS

Approval from the PhD Committee, School of Medicine, Faculty of Health Sciences, University of Pretoria was received on the 1st of November 2019. Approval from the Ethics Committee, Faculty of Health Sciences, University of Pretoria was received on the 22nd of January 2020 under protocol number 3/2020, and was renewed on an annual basis.

The custom designed TaqMan® OpenArray™ run on the Applied Biosystems QuantStudio 12K Flex Real-time PCR (Life Technologies, California, USA) was used to genotype the study population across the sixty pre-selected SNPs. The TaqMan® OpenArray™ was validated against the gold standard. The genotype and allele frequencies were calculated and compares to data from the 1000 Genomes Project extracted from HapMap data obtained from Ensembl. Two populations were assessed, namely European and African. Statistical analysis was conducted to compare the intra- and inter-population variability.

3.1   Study cohort

In order to have a sufficient number of DNA samples for the study, samples previously collected for various studies were included after obtaining approval from the Ethics Committee of the Faculty of Health Sciences of the University of Pretoria (protocol number: 3/2020, Appendix I). The study was carried out in accordance with the 2013 revised version of the Declaration of Helsinki.[100] The study included DNA samples from 813 self-identified black South African Bantu speakers. The samples included in the study were divided in four groups, according to how they were previously recruited. The first group (n = 192) included blood bank donors for the study of gene polymorphisms in the general population (Ethics protocol # 73/2006, Appendix II), the second group (n = 190) included women with breast cancer for the study of breast cancer genes (Ethics protocol #KCT265, Appendix III), the third group (n = 201) consisted of men with prostate cancer for the study of prostate cancer genes (43/2010, Appendix IV), and the last study included male Tshivenda individuals (n = 230) who were previously enrolled as controls in a study investigating epigenomic changes due

to pesticide exposure in a malaria area (Ethics protocol # 43/2003, Appendix V). Genetic analysis was core to all the above-mentioned projects and associated protocols, thus utilization of the DNA for further genetic analyses was within the scope of the initial ethics approval. Participants from the four study cohorts were all older than eighteen years of age. All samples were anonymized, prior to enrolment into the present study. Demographic information such as gender and ethnicity (home language) were available for 579 of the samples. The ethnic/linguistic groups are distinguished by common Bantu languages i.e. isiZulu, isiXhosa, Ndebelel, Swati, Tshivenda, Setswana, Sepedi and Sesotho. No additional medical history was required from participant. Participation in all four the studies were voluntary and participants were assured of their anonymity and confidentiality.

3.2 SNP selection

A bibliographic search was carried out to determine anticancer drug metabolic pathways and germline polymorphisms reported as being associated with drug efficacy and toxicity of cancer drugs. Ensembl was used to check allele frequencies across various population groups reported by various whole genome projects.[101] From a list of 100 single nucleotide variations, 60 were selected for the development of the TaqMan® OpenArray™ Real-Time PCR Plates (Applied Biosystems, Life Technologies, California, USA) (Table 6). The OpenArray™ format chosen for use in this study was the 64 assays and 48 samples per plate, which allowed for the genotyping of 960 samples. The selected SNPs with the flanking sequencing information was tabulated into an Excel spreadsheet and customization using the FileBuilder software (Applied BioSystems, Life Technologies, California, United States). Only one of the SNPs (rs1042028, SULT1A1*2), was a custom design probe based on the flanking sequence obtained from the NCBI SNP database. All other SNPs were tested and validated TaqMan® Assays (Applied Biosystems, Life Technologies, California, USA), purchased from the supplier.

Table 6. Selected single nucleotide polymorphisms associated with anticancer drug metabolism.[102]

Gene Name	Gene (star allele)	dbSNP	Location	Sequence [VIC/FAM]
ATP binding cassette subfamily B member 1	*ABCB1*	rs2032582	Chr.7:87531302	TATTTAGTTTGACTCACCTTCCCAG**[C/T]**ACCTTCTAGTTCTTTCTTATCTTTC
		rs1045642	Chr.7:87509329	TGTTGGCCTCCTTTGCTGCCCTCAC**[A/G]**ATCTCTTCCTGTGACACCACCCGGC
		rs1128503	Chr.7: 87550285	GCCCACTCTGCACCTTCAGGTTCAG**[A/G]**CCCTTCAAGATCTACCAGGACGAGT
ATP binding cassette subfamily C member 2	*ABCC2*	rs2273697	Chr.10: 99804058	CAACTTGGCCAGGAAGGAGTACACC**[A/G]**TTGGAGAAACAGTGAACCTGATGTC
		rs717620	Chr.10: 99782821	ACAATCATATTAATAGAAGAGTCTT**[C/T]**GTTCCAGACGCAGTCCAGGAATCAT
		rs3740066	Chr.10: 99844450	TCCTCAGAGGGATCACTTGTGACAT**[C/T]**GGTAGCATGGAGAAGGTAGGTGGAG
ATP binding cassette subfamily G member 1	*ABCG1*	rs225440	Chr.21: 42232943	CCATGTGGGTCAGATTAAATATATC**[T/C]**TGAAGGACTAAACCGTAAAACTAGG
ATP binding cassette subfamily G member 2 (Junior blood group)	*ABCG2*	rs2231142	Chr.4: 88131171	GCAAGCCGAAGAGCTGCTGAGAACT**[G/T]**TAAGTTTTCTCTCACCGTCAGAGTG
Carbonyl reductase 3	*CBR3*	rs8133052	Chr.21:36135203	GCTCCCCGCTCAGCCATGTCGTCCT**[A/G]**CAGCCGCGTGGCGCTGGTGACCGGG
		rs1056892	Chr.21:36146408	GGATGGGAAAGACAGCATCAGGACT**[A/G]**TGGAGGAGGGGGCTGAGACCCCTGT
Carbonyl reductase 1	*CBR1*	rs9024	Chr.21: 36073015	CTCTTATCAATTAGCACTCACTAAT**[A/G]**TACTACTAATTGAGCAACCTACGCA
Cytochrome P450 family 17 subfamily A member 1	*CYP17A1*	rs2486758	Chr.10: 102837723	GAGTCAAGGCTTGGAACTTTCCATG**[C/T]**TGCAAAATCAAAATCACTGGACAGA
Cytochrome P450 family 19 subfamily A member 1	*CYP19A1*	rs4646	Chr.15: 51210647	TCTGGTGTGAACAGGAGCAGATGAC**[A/C]**AATAGCACCTAGCTTGGTGACAACC
Cytochrome P450 family 1 subfamily A member 1	*CYP1A1*	rs2606345	Chr.15: 74724835	GCTCCCTGCAGTTGGCAATCTGTCA**[A/C]**CCTGATTGTCTCCCAGCAAAGGACA
	CYP1A1	rs1048943	Chr.15: 74720644	AAGAGAAAGACCTCCCAGCGGGCAA**[C/T]**GGTCTCACCGATACACTTCCGCTTG
Cytochrome P450 family 1 subfamily A member 2	*CYP1A2*	rs762551	Chr.15: 74749576	TGCTCAAAGGGTGAGCTCTGTGGGC**[C/A]**CAGGACGCATGGTAGATGGAGCTTA
Cytochrome P450 family 1 subfamily B member 1	*CYP1B1*	rs1056836	Chr.2: 38071060	AAGTTCTCCGGGGTTAGGCCACTTCA**[C/G]**TGGGTCATGATTCACAGACCACTGG

Gene Name	Gene (star allele)	dbSNP	Location	Sequence [VIC/FAM]
Cytochrome P450 family 2 subfamily B member 6	CYP2B6(*16)	rs28399499	Chr.19: 41012316	TGTACAGAGAGAGTCTACAGGGAGA[C/T]TGAACAGGTGATTGGCCCACATCGC
	CYP2B6(*5)	rs3211371	Chr.19: 41016810	CAAAATACCCCCAACATACCAGATC[C/T]GCTTCCTGCCCCGCTGAAGGGGCTG
	CYP2B6(*6)	rs3745274	Chr.19: 41006936	TCATGGACCCCACCTTCCTCTTCCA[G/T]TCCATTACCGCCAACATCATCTGCT
	CYP2B6(*8)	rs12721655	Chr.19: 41004377	CACTATGAGGGACTTCGGGATGGGA[A/G]AGCGGAGTGTGGAGGAGCGGATTCA
Cytochrome P450 family 2 subfamily C member 19	CYP2C19(*17)	rs12248560	Chr.10: 94761900	AAATTTGTGTCTTCTGTTCTCAAAG[C/T]ATCTCTGATGTAAGAGATAATGCGC
	CYP2C19(*2)	rs4244285	Chr.10: 94781859	TTCCCACTATCATTGATTATTTCCC[A/G]GGAACCCATAACAAATTACTTAAAA
	CYP2C19(*27)	rs7902257	Chr.10: 94761665	AGCTCTTCCTTCAGTTACACTGAGC[A/G]TTTCCCCTCTGCAGTGATGGAGAAG
Cytochrome P450 family 2 subfamily C member 8	CYP2C8(*2)	rs11572103	Chr.10: 95058349	TAATATCTTACCTGCTCCATTTTGA[A/T]CAGGAAGCAATCGATAAAGTCCCGA
	CYP2C8(*3)	rs10509681	Chr.10: 95038992	AAAGATATTTGGATTAGGAAATTCT[C/T]TGTCATCATGTAGCACGGAAGTCAG
	CYP2C8(*4)	rs1058930	Chr.10: 95058362	GCTCCATTTTGATCAGGAAGCAATC[C/G]ATAAAGTCCCGAGGATTGTTAACAT
Cytochrome P450 family 2 subfamily C member 9	CYP2C9(*8)	rs7900194	Chr.10: 94942309	GAGGACCGTGTTCAAGAGGAAGCCC[T/G]CTGCCTTGTGGAGGAGTTGAGAAAA
Cytochrome P450 family 2 subfamily D member 6	CYP2D6(*10)	rs1065852	Chr.22: 42130692	CCGGGCAGTGGCAGGGGGCCTGGTG[A/G]GTAGCGTGCAGCCCAGCGTTGGCGC
	CYP2D6(*17)	rs28371706	Chr.22: 42129770	ACGCGGCCCGAAACCCAGGATCTGG[G/A]TGATGGGCACAGGCGGGCGGTCGGC
	CYP2D6(*3)	rs35742686	Chr.22: 42128242	GGCTGGGCTGGGTCCCAGGTCATCC[T/-]GTGCTCAGTTAGCAGCTCATCCAGC
	CYP2D6(*4)	rs3892097	Chr.22: 42128945	AGACCGTTGGGGCGAAAGGGGCGTC[C/T]TGGGGGTGGGAGATGCGGGTAAGGG
Cytochrome P450 family 3 subfamily A member 4	CYP3A4(*15)	rs4986907	Chr.7: 99769804	GACAGGCTTGCCTGTCTCTGCTTCC[C/T]GCCTCAGATTTCTCACCAACACATC
	CYP3A4(*1B)	rs2740574	Chr.7: 99784473	TAAAATCTATTAAATCGCCTCTCTC[C/T]TGCCCTTGTCTCTATGGCTGTCCTC
	CYP3A4(*22)	rs35599367	Chr.7: 99768693	GTGCCAGTGATGCAGCTGGCCCTAC[G/A]CTGGGTGTGATGGAGACACTGAACT
Cytochrome P450 family 3 subfamily A member 5	CYP3A5(*3)	rs776746	Chr.7: 99672916	ATGTGGTCCAAACAGGGAAGAGATA[T/C]TGAAAGACAAAAGAGCTCTTTAAAG
	CYP3A5(*6)	rs10264272	Chr.7: 99665212	CTAAGAAACCAAATTTTAGGAACTT[C/T]TTAGTGCTCTCCACAAAGGGGTCTT
	CYP3A5(*7)	rs41303343	Chr.7: 99652770	CCATCTGTACCACGGCATCATAGGT[A/-]AGGTGGTGCCTGGAAGGAAAGAAAC
Dihydropyrimidine dehydrogenase	DPYD	rs115232898	Chr.1: 97699474	ACCAAAAAGAGCAATCTTTGCAGAA[C/T]AGGCTTCAGACATTTTTTCTGGGGG

43

Gene Name	Gene (star allele)	dbSNP	Location	Sequence [VIC/FAM]
Dihydropyrimidine dehydrogenase	DPYD	rs67376798	Chr.1:97082391	ACCACAGTTGATACACATTTCTTCA[A/T]CAATCATAGCCACAACTTGCTCTAC
	DPYD(*13)	rs55886062	Chr.1: 97515787	CCATCCAGCTTCAAAAGCTCTTCGA[A/C]TCATTGATGTGCTGGTGGCTGGAGT
	DPYD(*2A)	rs3918290	Chr.1: 97450058	TGTTTTAGATGTTAAATCACACTTA[C/T]GTTGTCTGGAAAGTCAGCCTTTAGT
Epoxide hydrolase 1	EPHX1	rs1051740	Chr.1: 225831932	GAAGCAGGTGGAGATTCTCAACAGA[C/T]ACCCTCACTTCAAGACTAAGATTGA
		rs2234922	Chr.1: 225838705	AAGCCCCCCCAGCTGCCCGCAGGCC[A/G]TACCCCGAAGCCCTTGCTGATGGTG
Excision repair cross complementation group 1	ERCC1	rs11615	Chr.19: 45420395	TTACGTCGCCAAATTCCCAGGGCAC[A/G]TTGCGCACGAACTTCAGTACGGGAT
		rs3212986	Chr.19:45409478	CACAGGCCGGGACAAGAAGCGGAAG[A/C]AGCAGCAGCAGCAGCCTGTGTAGTC
Hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1	HSD3B1	rs1047303	Chr.1: 119514623	TCCCTTGTGGACCGGCACAAGGAGA[C/A]CCTGAAGTCCAAGACTCAGTGATTT
Solute carrier family 22 member 1	SLC22A1	rs714368	Chr.6: 110456925	ATTGCCTGGGGGCCTGCAGACATGA[C/T]GAGGGGTGACTCCCATGAACACAGA
Solute carrier organic anion transporter family member 1B1	SLCO1B1	rs11045879	Chr.12: 21229685	TTCTTTGATGATATATATGAAGATG[C/T]TTGATTCTGTTATATTAACCCTGGA
Solute carrier organic anion transporter family member 1B3	SLCO1B3	rs4149117	Chr.12: 20858546	TATGGGAACTGGAAGTATTTTGACA[G/T]CTTTACCACATTTCTTCATGGGATA
		rs7311358	Chr.12: 20862826	CACTGGGATCTCTGTTTGCTAAAAT[A/G]TACGTGGATATTGGATATGTAGATC
		rs11045585	Chr.12: 20892760	TAGTAATCCTGAAGATTAAAGAAAC[A/G]TACTGACAGGGAAAATGGACTAGTA
Sulfotransferase 1A1	SULT1A1(*2)	rs1042028	Chr16:28606193	CTCCTGGGGGACGGTGGTGTAGTTGGTCATAGGGTTCTTCTTCATCTCCTTGAACGACGTGTGCTGAACCACGAAGTCCACGGTCTCCTTGGCAGGGAG[C/T]GCCCCACAAACTCCAGGATCTTTTGAATCTCCTTTTCGGGTTCTGAGCAGCAGAGGGCCCCTCAGTGGAGGCTCGGATTACTGATTCAGGAAAAGTAAA
Thiopurine methyltransferase	TPMT	rs1142345	Chr.6: 18130687	TCTCATTTACTTTTCTGTAAGTAGA[C/T]ATAACTTTTCAAAAAGACAGTCAAT
Testis-specific Y-encoded like proteine	TSPYL1	rs3828743	Chr.6: 116279647	ACGCCCCCCTCCTCTGAAGGCGGTG[A/G]AGGCGGGAGCGCGACGGTCTCCGAG
UDP glucuronosyltransferase family 1 member A6	UGT1A6(*2)	rs2070959	Chr.2: 233693545	GGGTTTTCCGTGTTCCCTGGAGCAT[A/G]CATTCAGCAGAAGCCCAGACCCTGT
	UGT1A6(*3)	rs1105879	Chr.2: 233693556	GTTCCCTGGAGCATACATTCAGCAG[A/C]AGCCCAGACCCTGTGTCCTACATTC
	UGT1A6(*4)	rs17863783	Chr.2: 233693631	ACCACATGACTTTTTCCCAACGAGT[G/T]GCCAACTTCCTTGTTAATTTGTTGG
UDP glucuronosyltransferase family 2 member B15	UGT2B15(*2)	rs1902023	Chr.4: 68670366	ATTTTCAGAAGAGAATCTTCCAAAT[A/C]ATTTTTAGTTAAAGATGTAGGATAA
X-ray repair cross complementing 1	XRCC1	rs25487	Chr.19: 43551574	GGGTTGGCGTGTGAGGCCTTACCTC[C/T]GGGAGGGCAGCCGCCGACGCATGCG

3.3 Genomic DNA extraction and quantification

Whole blood (10 mL) was collected from each participant using an BD Vacutainer®
EDTA-tube (Becton, Dickinson and Company, Franklin Lakes, USA). Genomic DNA
was extracted from the entire whole blood sample.

Genomic DNA extraction was carried out following the method of Johns and Paulus-
Thomas,[103] using sodium perchlorate. The extraction was performed in a 50 mL
polypropylene tube. The following was added to each tube; 10 mL whole blood and 30
mL cell lysis buffer (0.32 M sucrose/10 mM Tris-HCL/5 mM $MgCl2$/1% Triton X-100,
pH 8.0). Tubes were placed on ice for 10 min, followed by centrifugation for 20 min at
8120 g at 4°C (Beckman model J2-21M centrifuge, JA-17 rotor). The supernatant was
removed and discarded, and the pellet (leucocytes) was resuspended in 9 ml
suspension buffer (10mM Tris-HCl pH 8; 0.15M NaCl; 5mM EDTA; pH 8.0). For cell
membrane lysing, 1 mL of 10% sodium dodecyl sulfate (SDS) (BioUltra, Merck, South
Africa) was added to the tube. The content was adjusted to 1.5% SDS and incubated
at 60°C for 10 min in a water bath (in order to breakdown heat susceptible proteins).
Directly following incubation, 2.5 mL freshly prepared 5 M sodium perchlorate
($NaCLO4$) was added. The tube was inverted several times to mix the solution and
incubated on a rotating platform for 30 min at room temperature. Thereafter tubes were
centrifuged at 330 g for 15 min at room temperature (Beckman model J2-21M
centrifuge, JA-17 rotor), and the aqueous phase transferred to a new 50 mL
polypropylene tube, and the above process was repeated to remove excess proteins.
DNA was precipitated by the addition of 5 mL refrigerated absolute ethanol (molecular
grade). DNA was captured with a sterile siliconized glass rod and transferred to a new
15 mL polypropylene tube containing 0.5 mL Tris-EDTA resuspension buffer (10 mM
Tris-HCL/1 mM EDTA, pH 8.2).

Fluorescence based methods are preferred to spectrophotometric methods, gel-based
techniques, dye staining and blotting methods in the quantitation of human DNA. The
sensitivity of fluorescence-based detection is far greater, as only molecules that
fluoresce are detected, compared to UV absorption methods. Fluorescence assays
are generally quick to perform and inexpensive in comparison to other the traditional
methods.[104]

The purity and concentration of the DNA was checked using the Nanodrop (ND-2000, Thermo Fisher Scientific, Waltham, Massachusetts), while the concentration of intact double stranded DNA (dsDNA) was measured using the Qubit® 4.0 fluorometer (Q33226, Invitrogen, Applied BioSystems, Life Technologies, Carlsbad, USA) and the Qubit® Double Stranded DNA (dsDNA) Broad Range (BR) assay kit (Q32850, Invitrogen, Applied BioSystems, Life Technologies, Carlsbad, USA) according to manufacturer's instructions.[105] The Qubit® 4.0 is a stand-alone benchtop fluorometer (Figure 15) which is highly accurate, requires as little as 1 µL sample and produces results rapidly (within 5 min of start of analysis).[106] Prepared samples were aliquoted into clear 0.5 mL Qubit® assay tubes (Thermo Fisher Scientific, South Africa), vortexed and incubated in the dark for 2 min. The clear 0.5 mL Qubit® assay tubes were inserted into the Qubit® 4.0 one by one for analysis, selecting the appropriate dsDNA program and 1 µL input volume.
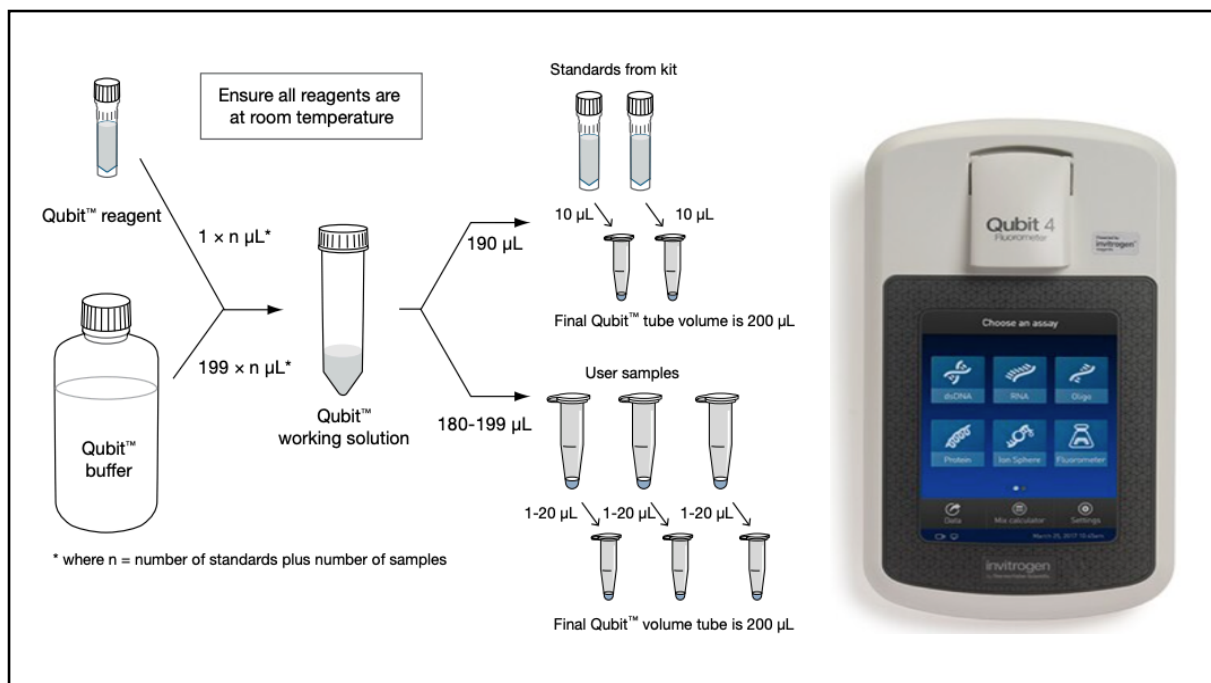
Figure 15. Qubit® 4.0 fluorometer, Invitrogen™ (Reproduced with permission[106])

3.3.1 DNA normalization and 96-well plate set-up

Once the concentration of the genomic DNA was determined, the solution dilution calculation (C1.V1 = C1.V2) was used to determine the dilution required to normalize all genomic DNA samples to a working stock concentration of 50 ng/µL, which is considered an acceptable starting concentration for the TaqMan® OpenArray™ genotyping platform.

To simplify the loading of the 384-well plate, 10 µL of the normalised genomic DNA sample was aliquoted into a well on the 96-well plate for which the format was specific for the OpenArray™ 64 SNP format loading strategy, as determined by the Sample Tracker Software. The 64-genotyping plate allows for 48 samples to be loaded onto a single OpenArray™. For each OpenArray™ plate, 46 samples were loaded, as well as one control sample with a known genotype (referred to as the positive control) and one non-template control (NTC) (Figure 16).

	1	2	3	4	5	6	7	8	9	10	11	12
A	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
B	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24
C	S25	S26	S27	S28	S29	S30	S31	S32	S33	S34	S35	S36
D	S37	S38	S39	S40	S41	S42	S43	S44	S45	S46	POS	NTC
E	S47	S48	S49	S50	S51	S52	S53	S54	S55	S56	S57	S58
F	S59	S60	S61	S62	S63	S64	S65	S66	S67	S68	S69	S70
G	S71	S72	S73	S74	S75	S76	S77	S78	S79	S80	S81	S82
H	S83	S84	S85	S86	S87	S88	S89	S90	S91	S92	POS	NTC

Figure 16. A schematic representation of a 96-well plate design, containing 92 samples, two positive controls (POS) and two non-template controls (NTC). Samples in row A-D were loaded onto one OpenArray™ plate, whereas samples in row E-H were loaded onto a second OpenArray™ plate.

3.4   The TaqMan® OpenArray™ genotyping

3.4.1 Principle

The TaqMan® OpenArray™ genotyping platform (Applied BioSystems, Life Technologies, California, USA) allows for high performance and high-throughput real-time polymerase chain reaction (PCR) technology. Furthermore, simultaneous analysis of several independent single nucleotide variations across several samples can occur. The default plate layout that was employed consisted of 60 single nucleotide variations for 48 samples.

The hydrolysis of the TaqMan® probes used for the TaqMan® OpenArray™ genotyping technique, requires a pair of primers, one associated with the wild-type sequence and the other with the mutant sequence, as well as two different Minor Groove Binder (MGB) probes for each assay. One probe for the normal or wild type sequence, is labelled with the VIC® fluorophore and the other probe for the mutant sequence, is labelled with the FAM® fluorophore (Figure 17). The plates used in the TaqMan® OpenArray™ assays were composed of 48 sub-arrays (4.5mm × 4.5mm), each with 64 nano-wells. The surface of the plate has hydrophobic properties, while the interior is hydrophilic in nature. The probes and primers required for the reactions are inserted in the interior of the wells by the manufacturer. These physical properties of the wells enable small quantities (33 nL) of sample to be loaded into the wells (Figure 18).[107,108]
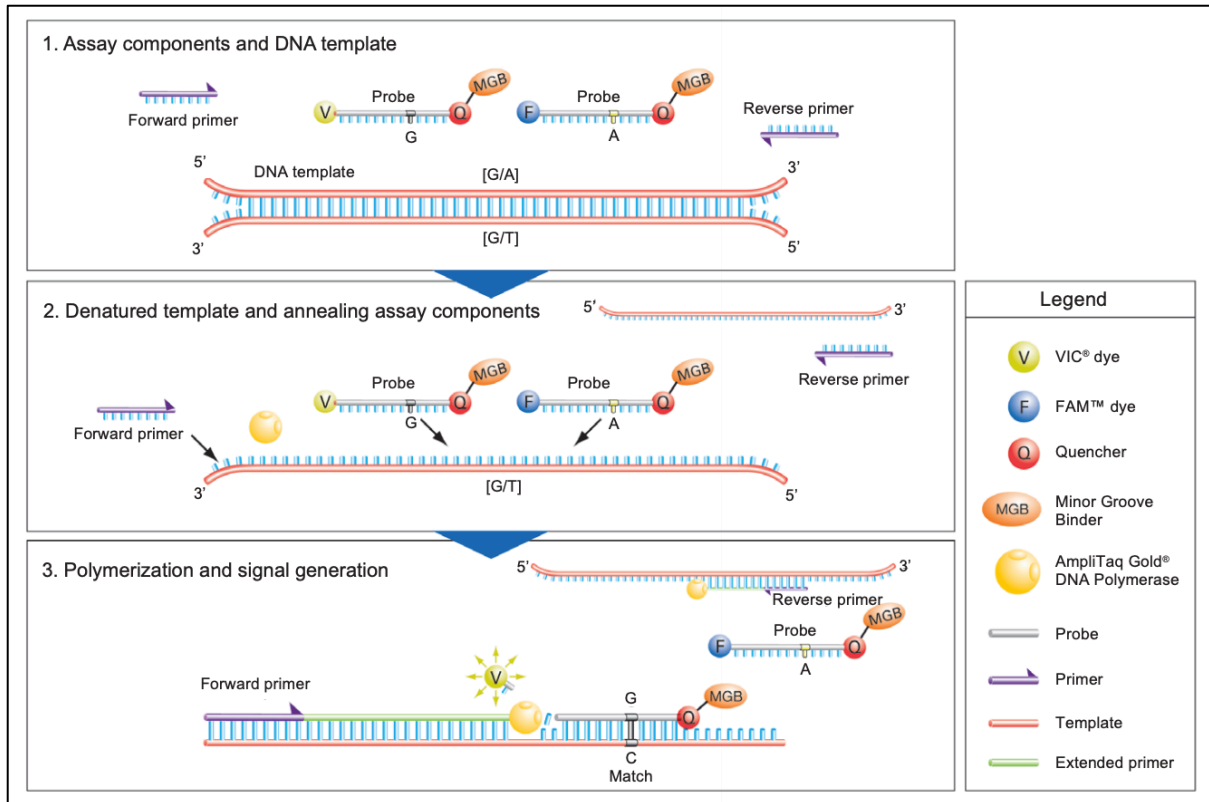
Figure 17. Selective annealing of TaqMan® MGB probes achieves allelic discrimination. (Reproduced with permission[108]).

An automated pipetting system (OpenArray™ AccuFill® System) (4471021, Applied BioSystems, Life Technologies, California, USA) is connected to the computer controls for loading of the DNA into the 384-well plate to the OpenArray™ by means of the manufacturer's proprietary software. The platform also includes the QuantStudio™ 12K Flex Real-Time PCR System (4471087, Applied BioSystems, Life Technologies, California, USA), suitable for cycling the OpenArray™ plates, a QuantStudio™ OpenArray™ Plate Press and an OpenArray™ block. The QuantStudio™ 12K Flex Real-Time PCR System is controlled by the QuantStudio™ 12K Flex Software (Applied BioSystems, Life Technologies, California, USA).
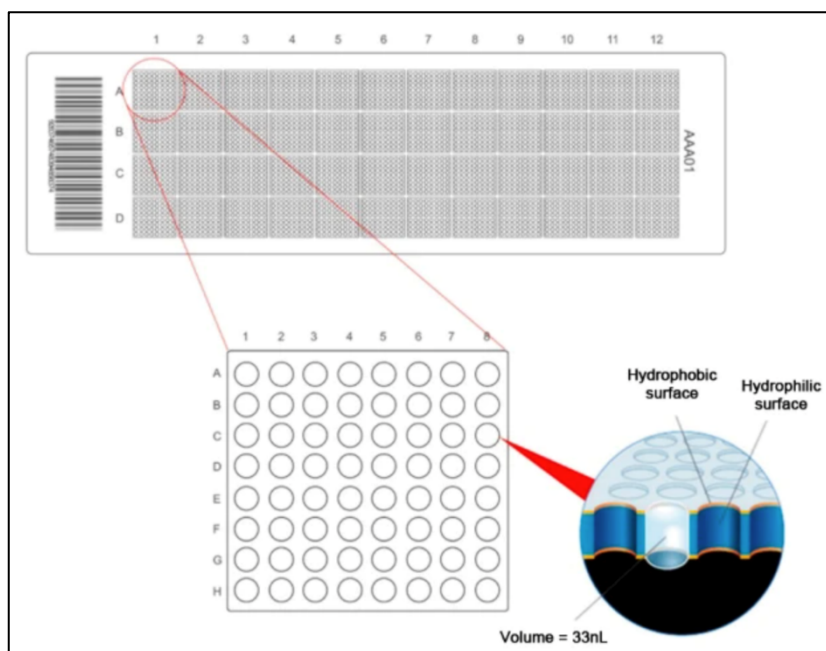
Figure 18. Representation of the OpenArray™ plate, containing 48 sub-arrays highlighting the arrangement of the sub-array with 64 nano-wells (Reproduced with permission[107]).

3.4.2 Assay design for OpenArray™ design

Primer design is the corner stone to any successful PCR reaction. There are more than a million predesigned TaqMan SNP genotyping human assays, which have been functionally tested and validated (Thermo Fisher Scientific). All 60 SNPs that were selected for the design of the 64 platform OpenArray™ plate were available, reducing the risk of assay failure. The assay makes use of the TaqMan® assay-based (5' nuclease) chemistry, enabling the amplification and detection of specific polymorphisms in the genomic DNA.

The TaqMan OpenArray™ format enabled the simultaneous detection of 60 genetic variations in 48 samples (46 study samples, one positive control and one NTC) in a single 3.5 h reaction. The Applied Biosystems QuantStudio 12K Flex instrument is capable of handling four OpenArray™ plates per run, thus enabling the capability of genotyping 192 samples per run (48 x 4).

3.4.3 Genotyping procedure

After the samples had been transferred to 96-well plates, 3 µL of each sample from the 96-well plate was transferred into a 384-well plate, containing 3 µL of the TaqMan® OpenArray™ Genotyping Master Mix. The distribution pattern of sample loading was determined by the Sample Tracker Software and the use of an adjustable pipette. A 12-channel adjustable pipette was utilized to transfer the samples from the 96-well plates to the 384-well plates. The 384-well plates were sealed with foil, gently shaken to ensure a uniform mixture of TaqMan® OpenArray™ Genotyping Master Mix and centrifuged at 12 000 g for 10 s to remove any bubbles. Bubbles present in the solution, can cause inaccurate loading by the automated pipetting system, causing some nanopores not to be filled with DNA-TaqMan master mix solution.

OpenArray™ plates were required to thaw for at least 30 min prior to loading, to ensure that the DNA-TaqMan master mix solution is securely inserted into the nanopores on the OpenArray™ plate.

The AccuFill® automated pipetting system was used to transfer the DNA from the 384-well plates to the TaqMan® OpenArray™ plates. A volume of 33 nL was dispensed into each nano-well. The TaqMan® OpenArray™ plate was sealed with a glass lid using the ™Plate Press and stabilized with immersion fluid. The OpenArray™ plate was sealed with a screw to prevent leaking of the immersion fluid during the reaction. The AccuFill® system creates a loaded file by merging the imported 384-well plate design file and the set-up file™ supplied by the manufacturer for each OpenArray™ plate. The OpenArray™ plate was then loaded into the QuantStudio™ 12K Flex Real-Time PCR System, using the QuantStudio™ software.[108] The QuantStudio™ software requires the importation of the loaded file. Standard thermocycling conditions were used: initial denaturation held at 95°C, followed by 40 cycles of 15 s denaturation at 92°C and 1 min annealing at 60°C.

In accordance with the TaqMan Genotyping user manual, the initial analysis was done using the QuantStudio™ software version 1.3 to check the overall quality of the run, whereafter the experiment was exported and imported into the TaqMan® Genotyper software version 1.6.0 for interpretation of the results according to the graphs of

clusters generated. Results were validated by the inclusion of non-template and positive controls (validated via sequencing).[108]

3.4.3.1 Genotyping analysis and quality control

Initial analysis was conducted using the QuantStudio™ software to evaluate the quality of the run. Quality control (QC) images were exported to assess the accuracy of the loading of the OpenArray™ plate. The experiment was analysed and saved after which the experiment file was imported into the TaqMan® Genotyper software. A text file containing the assay information which includes the information of the MGB probe assigned to each allele (supplied by the manufacturers) was also imported into the software. The software performed the analysis and interpretation of the results according to the graphs of clusters generated. The alleles are assigned to either the FAM or VIC probe, and the results are presented in a combination of either VIC/VIC, VIC/FAM or FAM/FAM. Samples falling outside of the three general clusters also referred to as outliners, required a more detailed analysis. The amplification of the probe can be tracked in real-time using the QuantStudio™ software to ensure accurate calling of the genotype. A more indepth explanation of the analysis and intrepretation of the results is provided in Appendix VI, which contains graphical illustrations of the analysis process.

3.4.3.2 Positive control sequencing

Sequencing is the gold standard for genetic genotyping. To ensure the TaqMan® OpenArray™ platform was working as expected, and to ensure the accuracy of the results, a positive control was included in each run. A NTC was also included into each run, to confirm the absence of reagent contamination and the potential of false positives. Amplification of the NTC, was indicative thereof that the run failed.

A control sample was selected to be the positive control for the test validation. Whole genome sequencing was performed on the positive control sample. The sequencing of the prepared whole genome library was performed at the South African Medical Research Council's (SAMRC) Genomic Center. Fragmentation was performed with

1000 ng DNA using the Covaris™ Focused ultrasonicator (Thermo Fisher Scientific, California, USA), to produce fragments of gDNA of between 100-700bp. Selection of 280 bp fragments was performed by magnetic bead-based size selection, using MGIEasy DNA Beads (1000005279, MGI Tech, Shenzhen, China). Library preparation was performed with 50 ng of fragmented DNA using the MGI Universal library prep kit, which includes MGIEasy DNA Adapters and a circularization module (1000006985, MGI Tech, Shenzhen, China), according to the manufacturer's instruction (Manual Version A1). The library preparation consists of three major steps, end repair and A-tailing, adapter ligation and PCR amplification. End-repair and A-tailing is an enzymatic reaction to convert fragmented DNA into 5'-phosphorlylated and 3'-dA-tailed DNA fragments, to enable ligation of the adapters. Once the adapters have been incorporated onto the DNA fragments, the fragments are amplified following seven PCR cycles. PCR cycles are limited, to reduce the potential of introducing high levels of duplications in the sequencing. The amplified adapter-ligated fragments were denatured and converted into single stranded rolling circles, using splint circularization. The single stranded circle DNA was thereafter converted into DNA nanoballs, which were loaded onto the sequencing PE100 flow cell. Pair-ended sequencing strategy was employed with a read length of 100 bp. Massively parallel sequencing was performed using DNA nanoball-based technology on the MGISEQ-2000RS (MGI Tech, Shenzhen, China), and the MGISEQ-2000RS High-throughput sequencing kit (PE100), (10000016950, MGI Tech, Shenzhen, China).

3.4.3.3 Reproducibility of genotype calls

The genotype of the positive control sample for the 60 SNPs was determined by Dr Brigitte Glanzmann (Bioinformatician at the SAMRC Genomics Center) using internal pre-established pipelines and aligning the sequence data to the human reference genome GRCh37. This control sample was thus used as the positive control and was included on each OpenArray™ plate that was processed. The genotyping results for the positive control for each plate was aligned to sequence data, to access the accuracy and reproducibility of the results.

3.5 Comparative database selection and data collection

The 1000 Genomes Project was initiated in 2008 and completed in 2015. It remains one of the most comprehensive genome databases, and global reference for human genetic variations. This global effort conduction sequencing of over 2500 genomes across 26 populations including Africa, East Asia, Europe, South Asia and the Americas. Participant samples were sequenced at low coverage and combined with array-based genotyping and supplemented with deep coverage exome sequencing. HapMap samples were also covered.[109] The database includes a broad spectrum of over 88 million genetic variations. The resource included >99% of SNP variants with a frequency of >1% for a variety of ancestries.[110] The Ensembl genome browser was utilised to extract allele frequencies of the 60 selected polymorphisms reported from the 1000 Genomes Project.[101] This data was exported to an Microsoft excel spreadsheet, for downstream analysis.

3.6 Data management and statistical analysis

The genotype and allele frequencies are the fundamental quantities of interest for polymorphic loci, as it measures genetic variation. Genotypic frequencies are referred to as the percentage of individuals in a population that have a specific genotype and show the distribution of genetic variation in the population. The allele frequency is the percentage of all copies of a certain gene in the population with a specific allele, and is the accurate measurement of the amount of genetic variation within a population.[111] The allele frequency can also be seen as the probability of a certain allele.

Genotype frequencies (P) are calculated by dividing the number of observations (N_{11}, N_{12}, N_{22}) by the population size (N). Loci with two alleles produce three possible genotypes (P_{11}, P_{12}, P_{22}). The total of the three genotype frequencies (P) must equal 1.[111]

$$P(genotype) = \frac{N(genotype)}{N}$$

$$P11 = \frac{N11}{N}$$

$$P12 = \frac{N12}{N}$$

$$P22 = \frac{N22}{N}$$

$$P11 + P12 + P22 = 1$$

For diploid individuals who carry two copies of each locus, 2N is designated. For simplicity reasons, allele frequency for allele 1 was designated *p* and allele 2 was designated *q* in the below calculation.

$$p = \frac{2N11 + N12}{2N}$$

$$q = \frac{2N22 + N12}{2N}$$

Therefore, the total of the two allele frequencies should equal 1 (*p* + *q* = 1). The formulas can be reshuffled to take into account the three potential genotype frequencies, *p* and *q* being represented as follows:

$$p = P11 + \frac{1}{2}P12$$

$$q = P22 + \frac{1}{2}P12$$

$$p2 + 2pq + q2 = 1$$

The allele frequency data for the populations of African and European ancestry recorded in the 1000 Genomes Project were obtained from the Ensembl website.[101]

Prof Michael Kidd from the Department of Statistics and Actuarial Sciences, University of Stellenbosch, assessed with the biostatistical analysis to ensure the accuracy of the analysis. The statistical and data science software, Stata, was used to perform the statistical analysis. The Hardy-Weinberg equilibrium (HWE) was used to determine the normality of the genotype and allele frequency distribution. Nucleotide homology was determined by performing a Blast search on probe sequences of SNPs that rejected the null hypothesis of HWE (*P*<0.05) and drawing a pair-wise alignment distance tree.

The Chi-squared and Fisher exact tests in conjunction with correspondence analysis (CA) was used to compare the observed allele frequencies, to the reported allele frequencies in the 1000 Genomes database per SNP. The observed allele frequencies of the selected genetic polymorphisms in the study population were compared to the reported allele frequencies for a population of African ancestry, followed by a population of European ancestry, to determine the variation between the European and other African populations. Principle component analysis (PCA) was performed on the minor allele frequencies to determine the variation between the study, African and European populations, using the programming language R.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Optimizing performance of the TaqMan® OpenArray™

Direct or targeted sequencing has been the gold standard for genotyping analysis; however, it has several draw backs in that it is expensive and time consuming. PCR based techniques are a cheaper option for the detection of main variations, however, they are largely individual and generate one result per reaction. Customisable, high-throughput, low-cost technologies have made it possible to perform simultaneous reactions in a single experiment, with low sample and reagent volume inputs required, increased accuracy, as well as simplicity.

The sodium perchlorate extraction method produced high quality DNA, with yields > 2 µg, and purity ratios between 1.8 and 2.0. The input volume for each sample was 3 µL, normalised to 50 ng/µL with TE buffer. Normalisation, although time consuming, is an essential step in the process as it improves downstream analysis, reflecting tighter clustering, and reduces potential for outliers.

Mastering the technique of loading the OpenArray™ plate was essential for consistent and good quality analysis. Quality control images were exported from the QuantStudio software, to assess the overall loading quality of the OpenArray™ plates, pre- and post-PCR reaction. Figure 19 below shows good and poor quality post-loading QC images of the OpenArray™ plates.

4.2 Validation of the TaqMan® OpenArray™

Accuracy of the OpenArray™, was confirmed by comparing the positive control's genotypic results obtained from the real-time PCR to the genotypic results determine through whole genome sequencing. The reproducibility of the OpenArray™ data was confirmed by the inclusion of the positive control on each OpenArray™ plate (Table 7). The OpenArray™ genotyping data were in agreement with the sequencing data of the positive control across all 17 plates analysed. A 99.9% amplification success with only one plate (GGQ02) that did not amplify the rs11572103 (*CYP2C8*) variant. None of the

OpenArray™ plates analysed showed amplification of the NTC, which is indicative of the absence of contamination.
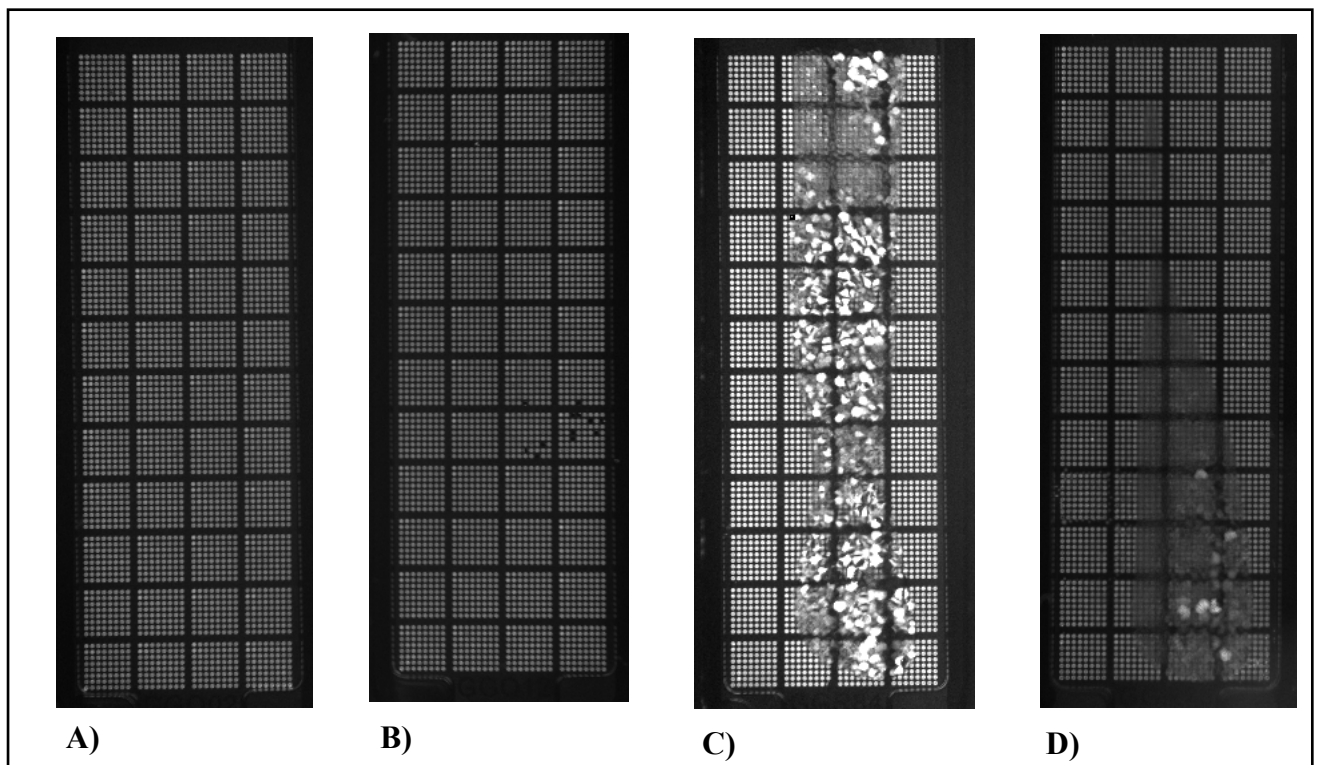
Figure 19. Quality control post-loading image captured reflecting the accuracy of loading of the OpenArray™ plates. A) A 100% accurately loaded plate, B) missing loading of nano-pores, C) Immersion oil leak on plate, D) poorly defrosted plate.

4.3 Genotyping analysis

The genotyping analysis is a two step process. The QuantStudio® 12K Flex Software version 1.3, was used to assess the overall quality of the run. The amplification of the samples are tracked in real-time and can be traced to ensure the correct genotypes are called by the TaqMan® Genotyper Software version 1.6.0. The TaqMan® Genotyper Software calls the genotypes of each sample for each individial SNP based on the assay information file imported and flourescent signal detected. If only the VIC flourescent probe is detected, the sample is classified as a homozygous for the VIC allele. Or alternatively homozygous for the FAM allele if only the FAM flourecence probe is detected. If both the VIC and FAM flourescent probes are detected, the

sample is classifed as a heterozygote for the respective alleles. Appendix VI contains more information and images on the genotypic analysis.

Overall the data provided distinct cluster groups and classification (homozygous for the VIC allele and homozygous for the FAM allele). Although distinct groupings were observed, some samples were borderline outliners. The multicomponent plots provided by the QuantStudio® 12K Flex Software, acted as a visional aid to show the amplification of the probes in real time. The custom design assay for *SULT1A1*2* (rs1042028), required more careful analysis, as the clustering between heterozygotes and homozygotes for the VIC allele, was not always clear for the two clusters (Figure 20). Genotype calling was made at the discretion of the analyser.

Table 7. Accuracy and reproducibility of OpenArray™ data confirmed through positive control genotyping data.

Gene name	dbSNP ID	WGS	OpenArray™ Barcode ID															
			GGR74	GGR75	GGQ02	GGQ04	GGQ05	GGQ12	GGQ10	GGQ11	GGR79	GGR80	GGQ06	GGQ07	GGQ09	GGQ39	GGQ40	GGQ41	GGR81
ABCB1	rs2032582	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
ABCB1	rs1045642	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
ABCB1	rs1128503	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
ABCC2	rs3740066	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
ABCC2	rs2273697	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
ABCC2	rs717620	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T
ABCG1	rs225440	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
ABCG2	rs2231142	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CBR1	rs9024	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CBR3	rs8133052	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
CBR3	rs1056892	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CYP17A1	rs2486758	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T
CYP19A1	rs4646	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
CYP1A1	rs1048943	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
CYP1A1	rs2606345	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
CYP1A2	rs762551	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
CYP1B1	rs1056836	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G
CYP2B6	rs12721655	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
CYP2B6	rs3211371	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
CYP2B6	rs28399499	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
CYP2B6	rs3745274	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T	G/T
CYP2C19	rs4244285	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G
CYP2C19	rs12248560	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T
CYP2C19	rs7902257	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CYP2C8	rs10509681	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T

Gene name	dbSNP ID	WGS	OpenArray™ Barcode ID																
			GGR74	GGR75	GGQ02	GGQ04	GGQ05	GGQ12	GGQ10	GGQ11	GGR79	GGR80	GGQ06	GGQ07	GGQ09	GGQ39	GGQ40	GGQ41	GGR81
CYP2C8	rs1058930	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G	C/G
CYP2C8	rs11572103	T/T	T/T	T/T	NO AMP	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
CYP2C9	rs7900194	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CYP2D6	rs1065852	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CYP2D6	rs3892097	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
CYP2D6	rs35742686	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
CYP2D6	rs28371706	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
CYP3A4	rs4986907	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
CYP3A4	rs2740574	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	G/G	T/T	T/T	T/T	T/T	G/G	T/T	T/T
CYP3A4	rs35599367	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	T/T	G/G	G/G	G/G	G/G	T/T	G/G	G/G	G/G
CYP3A5	rs776746	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
CYP3A5	rs10264272	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
CYP3A5	rs41303343	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
*DPYD*13*	rs55886062	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
DPYD	rs67376798	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
*DPYD*2A*	rs3918290	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
DPYD	rs115232898	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
EPHX1	rs2234922	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
EPHX1	rs1051740	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T	C/T
ERCC1	rs3212986	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C
ERCC1	rs11615	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G
HSD3B1	rs1047303	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
SLC22A16	rs714368	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
SLCO1B1	rs11045879	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
SLCO1B3	rs4149117	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
SLCO1B3	rs7311358	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A

Gene name	dbSNP ID	WGS	OpenArray™ Barcode ID																
			GGR74	GGR75	GGQ02	GGQ04	GGQ05	GGQ12	GGQ10	GGQ11	GGR79	GGR80	GGQ06	GGQ07	GGQ09	GGQ39	GGQ40	GGQ41	GGR81
SLCO1B3	rs11045585	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
SULT1A1	rs1042028	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
TPMT	rs1142345	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
TSPYL1	rs3828743	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
UGT1A6	rs2070959	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G	A/G
UGT1A6	rs17863783	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
UGT1A6	rs1105879	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C
UGT2B15	rs1902023	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C	A/C
XRCC1	rs25487	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
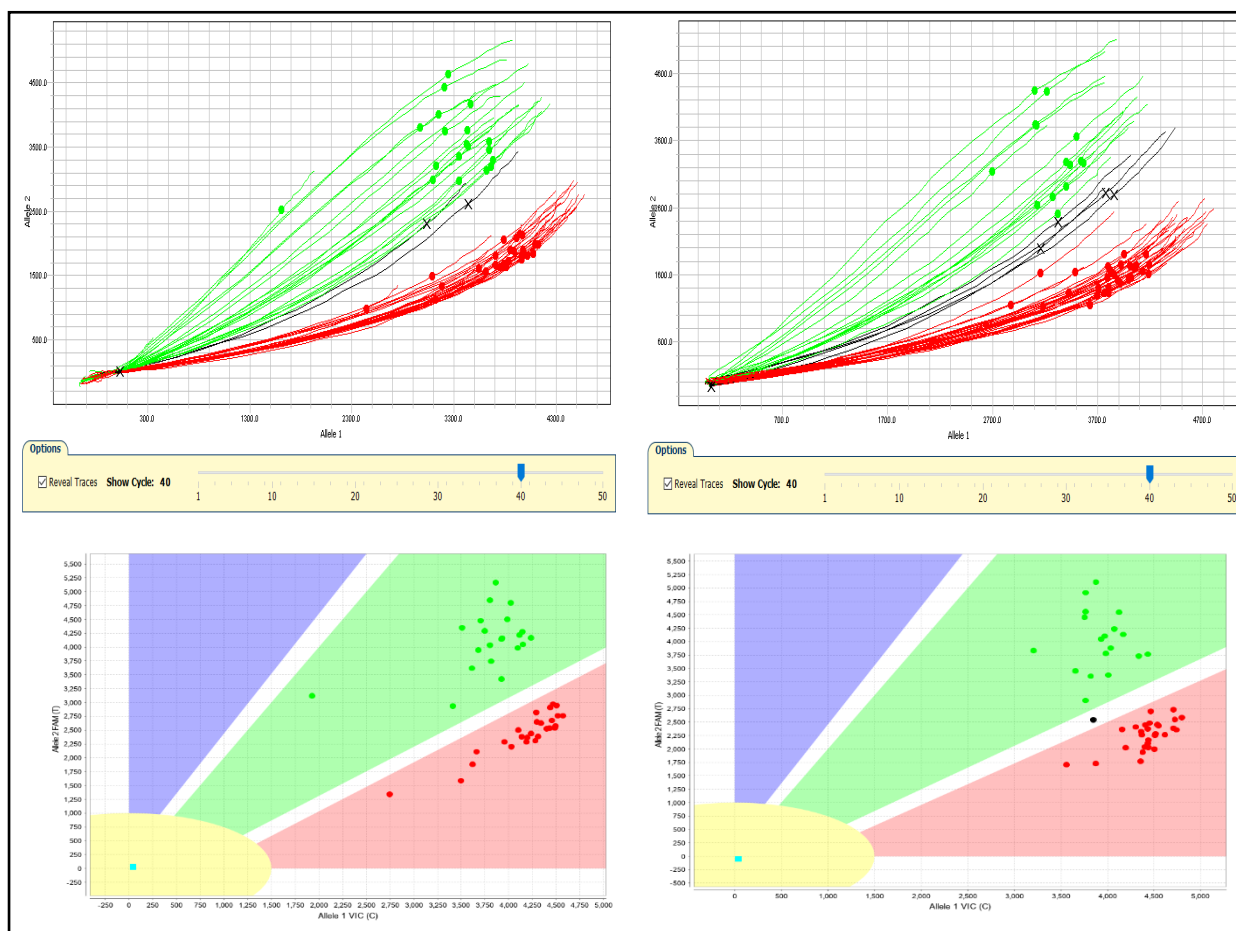
Figure 20. Comparison between QuantStudio® 12K Flex Software (top) and TaqMan® Genotyper Software (bottom). The classification tool on the TaqMan® Genotyper Software improves the clustering. The black dot in the TaqMan® Genotyper Software analysis in the bottom right, needs to be determined using the discretion of the analyst.

4.4 Cohort analysis

A total of 764 samples from the four cohorts of self-identified black South African Bantu speakers were analysed (Figure 21). The initial cohort consisted of 813 samples, however only 782 could be genotyped, due to the limited number of OpenArrays™ available. Customised arrays had a minimum order of 20 plates, and only one array would have been required to complete the analysis of the 813 samples. A further 18 samples were excluded from the analysis, these samples showed poor amplification in more than 10% of the SNPs assessed (Figure 21). Of the 764 samples, 263 were female (34.4%), and 501 were male (65.6%).

63
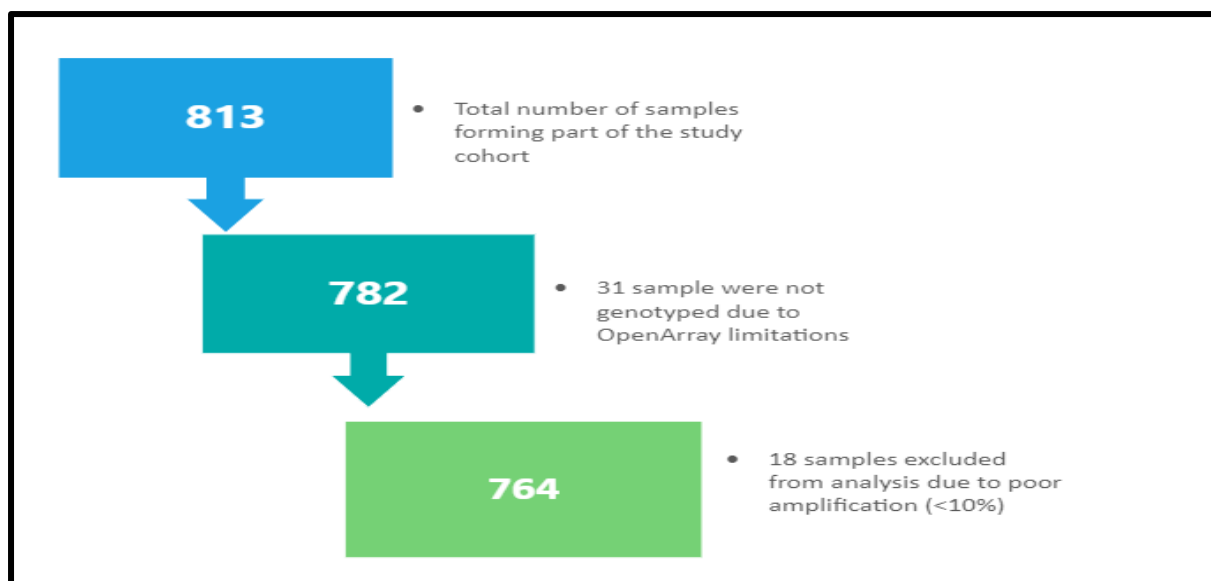
Figure 21. Total number of samples in the cohort, and the exclusion of 49 samples from the final analysis.

All participants in the study, were over the age of 18 years and self-identified themselves as black South Africans, speaking one of the twelve Bantu lanuages, nine of which formed part of the eleven official languages of South Africa. Three of the participants classified themselves as Afrikaans speakers of mixed ancestry.

Nearly a quarter of the study population (24%), classifed themselves as Bantu speakers and did not make mention their exact mother tongue and therefore did not form part of the intrapopulation assessment. The largest portion of the study group were TshiVenda speakers (30.2%). The four major ethnic division amoung Black South Africans are the Nguni, Sotho-Tswana, Shangaan-Tsonga and Venda. The major Sotho groups are the South Sotho (Basuto and Sotho), the West Sotho (Tswana), and the North Sotho (Pedi). Nguni languages include Xhosa, Zulu, Ndebele (sometimes referred to as "Northern Ndebele") and Swati (Table 8). The Sotho subfamily occupy mostly the Free-State, North-West province. The TshiVenda speakers occupy northern Limpopo around the border of Zimbabwe. While the Nguni-Tsonga speakers occupy most parts of the Eastern Cape, Kwa-Zulu Natal, Limpopo and Mpumalanga (Figure 22).

Table 8. Linguistic grouping distribution.

Language	Subfamily	Numbers
Afrikaans	Afrikaans	3
English	English	0
IsiNdebele	Nguni	16
IsiXhosa	(n = 107)	10
IsiZulu		70
SiSwati		11
Shangaan	Shangaan – Tsonga	1
Xitsonga	(n = 19)	18
Sesotho	Sotho-Tswana	38
Sesotho sa Leboa	(n = 219)	40
Setswana		141
TshiVenda	Venda (n = 231)	231
Bantu speakers	Unclassified	185
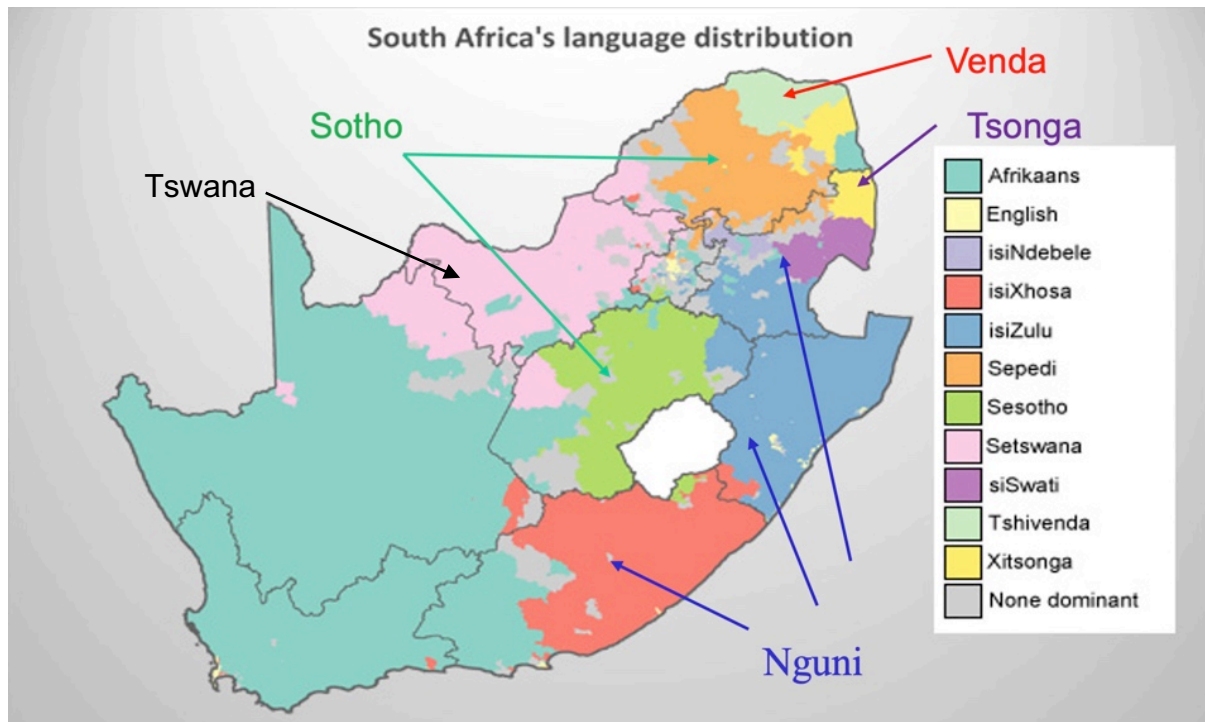
Figure 22. Geographical distribution of the official South African languages. (Reproduced with permission[112]).

4.5 Genotype and allele frequencies

The genotype and allele frequencies are summarized in Tables 9 and 10, respectively. Six of the sixty SNPs showed 100% amplification across all the samples analysed, namely, rs1272155 (*CYP2B6*), rs11572103 (*CYP2C8*), rs35599367 (*CYP3A4*), rs55886062 (*DPYD*), rs714368 (*SLC22A16*) and rs1902023 (*UGT2B15*). The four SNPs from the *CYP2D6* locus (rs1065852, rs3892097, rs35742686 and rs28371706) showed the poorest amplification, frequency of no amplification > 0.01 (Table 9). Previous studies have shown the CYP2D6 gene region to be highly polymorphic, with gene duplications and deletions also being reported, as well as hybrid genes.[59] Amplification of the *CYP2D6* variants ranged from 97.6 to 98.6% of the samples. The variation seen in the amplification could be the results of non-specific binding; gene deletion; or varied nucleotide changes due to novel genotypes. The exact cause of this variation could not be determined in the present study.

Five SNPs showed no genetic variability, with only homozygotes for the ancestral allele being detected in the study population namely, rs2032582 - C *(ABCB1),* rs12721655 - A (*CYP2B6*), rs7900194 – G (*CYP2C9*), rs55886062 - A and rs67376798 - T (*DPYD*). These alleles appeared to be fixed in the population, likely because of genetic drift and absence of gene flow. One *DPYD* variant has been identified as being African specific (rs115232898 – C) and has previously been reported to be common in Africa (1 - 4%). A low allele frequency of 2% was found in the present study population, which is in agreement with a previous study.[95] This variant is known to reduce the function of the dihydropyrimidine dehydrogenase enzyme and has been linked to severe toxicity when treated with 5-FU, doxorubicin and cyclophosphamide (FAC) chemotherapy protocol or Tegafur.[113]

Although rs7900194 (*CYP2C9)* was only detected with the ancestral allele, this SNP is multi-allelic (G/A/C/T), with the A allele presenting in 5% of the African population (Figure 23). The TaqMan assay used in the present study was designed specific for the G > T alleles. The manner in which the probe is designed, would result in G/A genotypes presenting as G/G genotypes during TaqMan genotyping analysis, which means that only the VIC probe would amplify. A/A genotypes would possibly present as samples failing to amplify, and in total seven samples showed a lack of

amplification. Similarly, rs2032582 *(ABCB1)* was also tri-allelic (A/C/T), with the T not being present in the African population (Figure 24). The assay design is specific for the C > T alleles. Only the ancestral C allele was observed in the study population. As the assay was not designed to detect tri-allelic polymorphism, the presence of the A allele cannot be excluded. From this data it could be confirmed that the T allele for rs7900194 and rs2032582 does not appear to be prevalent in the SSA population. Caution needs to be applied, as no definite conclusion can be drawn from the above-mentioned SNPs. Genotyping assays are designed to detect typical biallelic SNPs, yet an increasing number of pharmaceutical relevant SNPs of clinical value are known to be tri-allelic, meaning there are three different nucleotide bases that occur at the same loci location in the human population. Although tri-allelic SNPs can be detected using two paired conventional TaqMan assays, the OpenArray™ platform does not allow for this investigation.
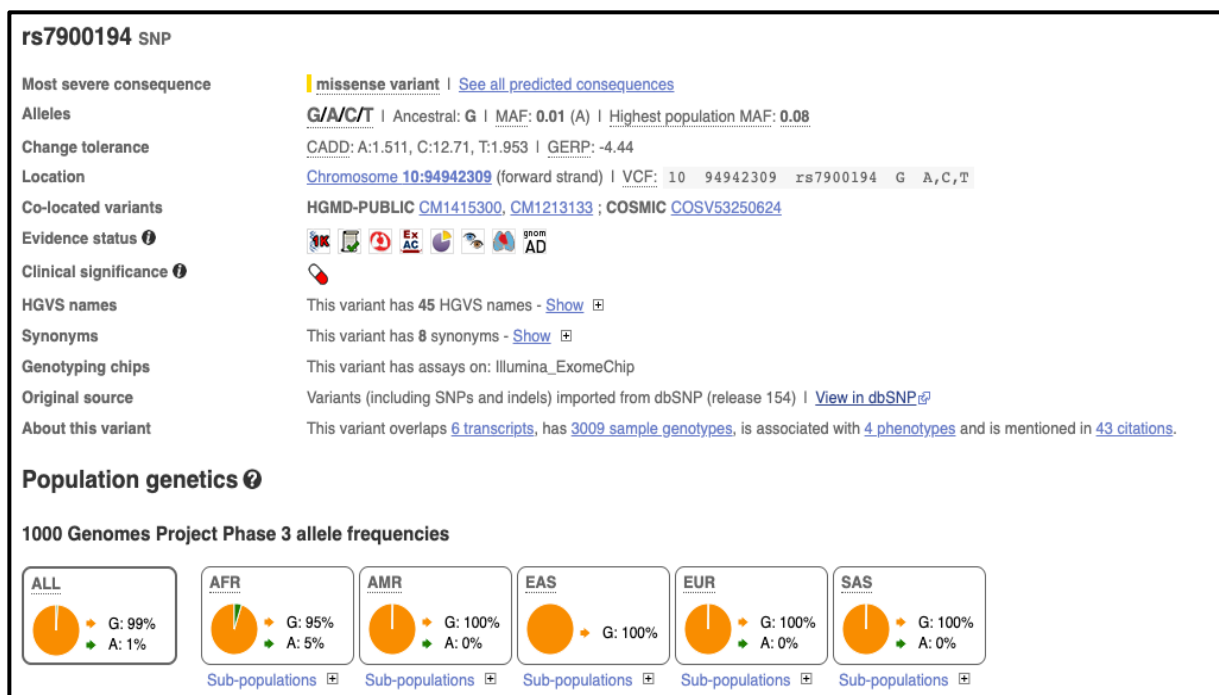
Figure 23. Ensembl population genetics summary of rs7900194 (*CYP2C9*). (Reproduced with permission[114])
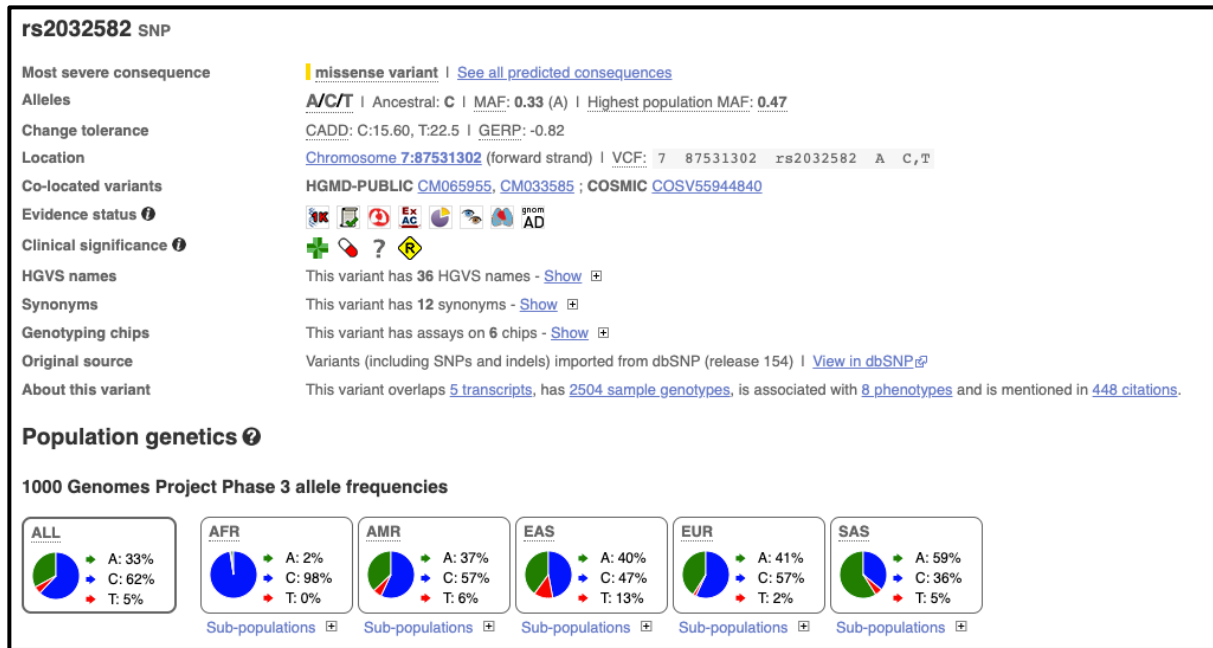
Figure 24. Ensembl population genetics summary of rs2032582 (*ABCB1*). (Reproduced with permission[115])

Genotyping data sets may contain errors which may lead to false conclusions. It is common practice to check whether observed genotypes conform to Hardy-Weinberg expectations. The HWE was first described in the early part of the twentieth century shortly after Mendel's work of inheritance patterns, and has become a corner stone in the history of population genetics.[116,117] The HWE makes the assumptions; that mating occurs randomly and that natural selection, migration, mutation and genetic drift is absent. The statistical power of any dataset, is strongly linked to the size of the population.[111] Deviation from HWE in random samples may be indicative of problematic assays; either due to non-specificity of assays or genotyping errors.[118]

Six of the selected SNPs, rs1065852, rs3892097, rs35742686 and rs28371706 from *CYP2D6* family*,* rs11615 (*ERCC1*) and rs1042028 (*SULT1A1*) were found to deviate from HWE ($P < 0.05$), indicating statistical significance and rejection of the null hypothesis (Table 11). SNP ascertainment bias which is the systematic deviation of population genetic statics from theoretical expectations, is the result of either nonrandom sampling or biased SNP discovery protocols. The customized array was created to assess a set of pre-ascertained SNPs of clinical relevance within the South African population. A statistical study conducted by Lachance and Tishkoff, showed

that genotyping arrays cause allele frequency distribution to shift toward the intermediate frequency allele, as a result SNP selection bias.[119] Sample size also play a contributing factor as common alleles are more likely to be found in smaller samples sizes.

Other possible explanations for the SNPs that showed deviation from HWE were explored. Technical reasons, such as assay non-specificity can impact the distribution of the genotype for any one variant, making any association unreliable. To confirm or exclude the possibility of non-specificity of the probe in the assay, a blast search was performed on the flanking sequencing of each of the six SNPs. Probe sets showing 100% homology with multiple regions in the human genome are considered "non-specific".[118] The flanking sequencing for each probe was obtained from the National Center for Biotechnology Information (NCBI) SNP database (Table 12).[102] Homology searches conducted using the NCBI basic local alignment search tool (BLAST)[120] and distance trees reflecting the homology are provided in Figures 39 to 44, Appendix VII. The flanking sequences were copied into the NCBI BLAST tool and compared to *Homo sapiens* (taxid: 9606).

CYP2D6 is involved in the metabolism of > 25% of all registered drugs. More than 120 allelic variants of *CYP2D6* have been identified and the phenotypic effect of many of these are well characterised. Dosing guidelines have also been developed by the CPIC, for many of the drug substrates. Previous studies have found the non-functional CYP2D6 allele *3 and decreased function allele *17 to be more prevalent in the African population,[59] thus making them important SNPs to monitor. The *ERCC1* gene is involved in the excision repair pathway and is required for the repair of DNA lesions. Mutations in this gene have been linked to toxicity in BrCa patients treated with 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy, commonly referred to as the FAC protocol.[121] The *SULT1A1* gene encodes for the sulfotransferase enzyme that catalyses the sulphate conjugate of many hormones, neurotransmitters and drugs. *SULT1A1*2* is involved in the elimination of the active metabolites of tamoxifen. Studies that investigated polymorphisms have shown a two-fold lower sulphation of the antiestrogenic metabolite, 4-hydroxy-tamoxifen.[45]

Table 9. Genotype frequency of the 60 SNPs analysed in the study population.

N = 764		Homozygous (VIC/VIC)		Heterozygous (VIC/FAM)		Homozygous (FAM/FAM)				
dbSNP	Gene (star allele)	Count	Frequency	Count	Frequency	Count	Frequency	Total	No Amplification	Amplification Success (%)
rs2032582	ABCB1	CC (n) = 761	**1,000**	CT (n) = 0	0,000	TT (n) = 0	0,000	761	3	99,61
rs1045642	ABCB1	AA (n) = 9	0,012	AG (n) = 143	0,187	GG (n) = 611	0,801	763	1	99,87
rs1128503	ABCB1	AA (n) = 5	0,007	AG (n) = 140	0,183	GG (n) = 618	0,810	763	1	99,87
rs3740066	ABCC2	CC (n) = 590	0,774	CT (n) = 159	0,209	TT (n) = 13	0,017	762	2	99,74
rs2273697	ABCC2	AA (n) = 28	0,037	AG (n) = 197	0,259	GG (n) = 535	0,704	760	4	99,48
rs717620	ABCC2	CC (n) = 740	0,974	CT (n) = 20	0,026	TT (n) = 0	0,000	760	4	99,48
rs225440	ABCG1	TT (n) = 276	0,363	TC (n) = 357	0,470	CC (n) = 127	0,167	760	4	99,48
rs2231142	ABCG2	GG (n) = 756	0,992	GT (n) = 6	0,008	TT (n) = 0	0,000	762	2	99,74
rs9024	CBR1	AA (n) = 0	0,000	AG (n) = 7	0,009	GG (n) = 754	0,991	761	3	99,61
rs8133052	CBR3	AA (n) = 67	0,088	AG (n) = 306	0,402	GG (n) = 389	0,510	762	2	99,74
rs1056892	CBR3	AA (n) = 172	0,226	AG (n) = 388	0,510	GG (n) = 201	0,264	761	3	99,61
rs2486758	CYP17A1	CC (n) = 1	0,001	CT (n) = 46	0,060	TT (n) = 715	0,938	762	2	99,74
rs4646	CYP19A1	AA (n) = 126	0,165	AC (n) = 351	0,461	CC (n) = 285	0,374	762	2	99,74
rs1048943	CYP1A1	CC (n) = 0	0,000	CT (n) =1	0,001	TT (n) = 762	0,999	763	1	99,87
rs2606345	CYP1A1	AA (n) = 1	0,001	AC (n) = 11	0,014	CC (n) = 749	0,984	761	3	99,61
rs762551	CYP1A2	CC (n) 142	0,186	CA (n) = 378	0,496	AA (n) = 242	0,318	762	2	99,74
rs1056836	CYP1B1	CC (n) = 497	0,654	CG (n) 226	0,297	GG (n) = 37	0,049	760	4	99,48
rs12721655	CYP2B6(*8)	AA (n) = 764	**1,000**	AG (n) = 0	0,000	GG (n) = 0	0,000	764	0	**100**
rs3211371	CYP2B6(*5)	CC (n) = 753	0,989	CT (n) = 8	0,011	TT (n) = 0	0,000	761	3	99,61
rs28399499	CYP2B6(*16)	CC (n) = 12	0,016	CT (n) = 152	0,201	TT (n) = 594	0,784	758	6	99,21
rs3745274	CYP2B6(*6)	GG (n) = 311	0,408	GT (n) = 358	0,470	TT (n) = 93	0,122	762	2	99,74
rs4244285	CYP2C19(*2)	AA (n) = 25	0,033	AG (n) = 206	0,271	GG (n) = 530	0,696	761	3	99,61

N = 764		Homozygous (VIC/VIC)		Heterozygous (VIC/FAM)		Homozygous (FAM/FAM)			No	Amplification
dbSNP	Gene (star allele)	Count	Frequency	Count	Frequency	Count	Frequency	Total	Amplification	Success (%)
rs12248560	CYP2C1(*17)	CC (n) = 536	0,704	CT (n) = 207	0,272	TT (n) = 18	0,024	761	3	99,61
rs7902257	CYP2C19(*27)	AA (n) = 18	0,024	AG (n) = 193	0,254	GG (n) = 550	0,723	761	3	99,61
rs10509681	CYP2C8(*3)	CC (n) = 0	0,000	CT (n) = 6	0,008	TT (n) = 754	0,992	760	4	99,48
rs1058930	CYP2C8(*4))	CC (n) = 0	0,000	CG (n) = 12	0,016	GG (n) = 749	0,984	761	3	99,61
rs11572103	CYP2C8(*2)	AA (n) = 31	0,041	AT (n) = 257	0,336	TT (n) = 476	0,623	764	0	**100**
rs7900194	CYP2C9(*8)	TT (n) = 0	0,000	TG (n) = 0	0,000	GG (n) = 757	**1,000**	757	7	99,08
rs1065852	CYP2D6(*10)	AA (n) = 28	0,037	AG (n) = 88	0,118	GG (n) = 632	0,845	748	16	**97,91**
rs3892097	CYP2D6(*4)	CC (n) = 713	0,947	CT (n) = 35	0,046	TT (n) = 5	0,007	753	11	**98,56**
rs35742686	CYP2D6(*3)	TT (n) = 749	0,999	T/- (n) = 0	0,000	-/- (n) = 1	0,001	750	14	**98,17**
rs28371706	CYP2D6(*17)	GG (n) = 477	0,640	GA (n) = 198	0,266	AA (n) = 70	0,094	745	19	**97,51**
rs4986907	CYP3A4(*15)	CC (n) = 730	0,964	CT (n) = 26	0,034	TT (n) = 1	0,001	757	7	99,08
rs2740574	CYP3A4(*1B)	CC (n) = 432	0,566	CT (n) = 285	0,374	TT (n) = 46	0,060	763	1	99,87
rs35599367	CYP3A4(*22)	GG (n) = 761	0,996	GA (n) = 3	0,004	AA (n) = 0	0,000	764	0	**100**
rs776746	CYP3A5(*3)	TT (n) = 539	0,708	CT (n) = 202	0,265	CC (n) = 20	0,026	761	3	99,61
rs10264272	CYP3A5(*6)	CC (n) = 488	0,642	CT (n) = 240	0,316	TT (n) = 32	0,042	760	4	99,48
rs41303343	CYP3A5(*7)	AA (n) = 7	0,009	A/- (n) = 142	0,187	-/- (n) = 612	0,804	761	3	99,61
rs55886062	DPYD*13	AA (n) = 764	**1,000**	AC (n) = 0	0,000	CC (n) = 0	0,000	764	0	**100**
rs67376798	DPYD	AA (n) = 0	0,000	AT (n) = 0	0,000	TT (n) = 762	**1,000**	762	2	99,74
rs3918290	DPYD*2A	CC (n) = 758	0,999	CT (n) = 1	0,001	TT (n) = 0	0,000	759	5	99,35
rs115232898	DPYD	CC (n) =1	0,001	CT (n) = 25	0,033	TT (n) = 737	0,966	763	1	99,87
rs2234922	EPHX1	AA (n) = 398	0,522	AG (n) = 305	0,400	GG (n) = 60	0,079	763	1	99,87
rs1051740	EPHX1	CC (n) = 42	0,055	CT (n) = 288	0,379	TT (n) = 430	0,566	760	4	99,48
rs3212986	ERCC1	AA (n) = 41	0,054	AC (n) = 275	0,360	CC (n) = 447	0,586	763	1	99,87
rs11615	ERCC1	AA (n) = 7	0,009	AG (n) = 64	0,084	GG (n) = 691	0,907	762	2	99,74
rs1047303	HSD3B1	CC (n) = 14	0,018	CA (n) = 133	0,175	AA (n) = 612	0,806	759	5	99,35
rs714368	SLC22A16	CC (n) = 145	0,190	CT (n) = 389	0,509	TT (n) = 230	0,301	764	0	**100**
rs11045879	SLCO1B1	CC (n) = 5	0,007	CT (n) = 90	0,118	TT (n) = 668	0,875	763	1	99,87

N = 764		Homozygous (VIC/VIC)		Heterozygous (VIC/FAM)		Homozygous (FAM/FAM)			No	Amplification
dbSNP	Gene (star allele)	Count	Frequency	Count	Frequency	Count	Frequency	Total	Amplification	Success (%)
rs4149117	*SLCO1B3*	GG (n) = 58	0,076	GT (n) = 317	0,417	TT (n) = 385	0,507	760	4	99,48
rs7311358	*SLCO1B3*	AA (n) = 60	0,079	AG (n) = 317	0,415	GG (n) = 386	0,506	763	1	99,87
rs11045585	*SLCO1B3*	AA (n) = 550	0,722	AG (n) = 193	0,253	GG (n) = 19	0,025	762	2	99,74
rs1042028	*SULT1A1(*2)*	CC (n) = 415	0,545	CT (n) = 346	0,454	TT (n) = 1	0,001	762	2	99,74
rs1142345	*TPMT*	CC (n) = 2	0,003	CT (n) = 63	0,083	TT (n) = 698	0,915	763	1	99,87
rs3828743	*TSPYL1*	AA (n) = 2	0,003	AG (n) = 58	0,076	GG (n) = 703	0,921	763	1	99,87
rs2070959	*UGT1A6*	AA (n) = 551	0,722	AG (n) = 198	0,260	GG (n) = 14	0,018	763	1	99,87
rs17863783	*UGT1A6*	GG (n) = 573	0,752	GT (n) = 173	0,227	TT (n) = 16	0,021	762	2	99,74
rs1105879	*UGT1A6*	AA (n) = 476	0,624	AC (n) = 258	0,338	CC (n) = 29	0,038	763	1	99,87
rs1902023	*UGT2B15*	AA (n) = 84	0,110	AC (n) = 309	0,404	CC (n) = 371	0,486	764	0	**100**
rs25487	*XRCC1*	CC (n) = 607	0,799	CT (n) = 143	0,188	TT (n) = 10	0,013	760	4	99,48

* Blue numbers indicate SNPs that showed 100% amplification across all the samples
* Red numbers indicate the SNPs that showed no genetic variation and the presence of a single genotype within the population
* Green numbers indicate the CYP2D6 polymorphisms showing amplification success < 99%

Table 10. The allele frequencies of the 60 SNPs analysed in the study population.

dbSNP	Gene (star allele)	VIC Allele	Count	Frequency	FAM Allele	Count	Frequency
rs2032582	ABCB1	C	1522	1,000	T	0	**0,000**
rs1045642	ABCB1	A	161	0,106	G	1365	0,894
rs1128503	ABCB1	A	150	0,098	G	1376	0,902
rs3740066	ABCC2	C	1339	0,879	T	185	0,121
rs2273697	ABCC2	A	253	0,166	G	1267	0,834
rs717620	ABCC2	C	1500	0,987	T	20	0,013
rs225440	ABCG1	T	909	0,598	C	611	0,402
rs2231142	ABCG2	G	1518	0,996	T	6	0,004
rs9024	CBR1	A	7	0,005	G	1515	0,995
rs8133052	CBR3	A	440	0,289	G	1084	0,711
rs1056892	CBR3	A	732	0,481	G	790	0,519
rs2486758	CYP17A1	C	48	0,031	T	1476	0,969
rs4646	CYP19A1	A	603	0,396	C	921	0,604
rs1048943	CYP1A1	C	1	0,001	T	1525	0,999
rs2606345	CYP1A1	A	13	0,009	C	1509	0,991
rs762551	CYP1A2	C	662	0,434	A	862	0,566
rs1056836	CYP1B1	C	1220	0,803	G	300	0,197
rs12721655	CYP2B6(*8)	A	1528	1,000	G	0	**0,000**
rs3211371	CYP2B6(*5)	C	1514	0,995	T	8	0,005
rs28399499	CYP2B6(*16)	C	176	0,116	T	1340	0,884
rs3745274	CYP2B6(*6)	G	980	0,643	T	544	0,357
rs4244285	CYP2C19(*2)	A	256	0,168	G	1266	0,832
rs12248560	CYP2C19(*17)	C	1279	0,840	T	243	0,160
rs7902257	CYP2C19(*27)	A	229	0,150	G	1293	0,850
rs10509681	CYP2C8(*3)	C	6	0,004	T	1514	0,996
rs1058930	CYP2C8(*4)	C	12	0,008	G	1510	0,992
rs11572103	CYP2C8(*2)	A	319	0,209	T	1209	0,791
rs7900194	CYP2C9(*8)	T	0	**0,000**	G	1514	1,000
rs1065852	CYP2D6(*10)	A	144	0,096	G	1352	0,904
rs3892097	CYP2D6(*4)	C	1461	0,970	T	45	0,030
rs35742686	CYP2D6(*3)	T	1498	0,999	-	2	0,001
rs28371706	CYP2D6(*17)	G	1152	0,773	A	338	0,227
rs4986907	CYP3A4(*15)	C	1486	0,982	T	28	0,018
rs2740574	CYP3A4(*1B)	C	1149	0,753	T	377	0,247
rs35599367	CYP3A4(*22)	G	1525	0,998	A	3	0,002
rs776746	CYP3A5(*3)	T	1280	0,841	C	242	0,159
rs10264272	CYP3A5(*6)	C	1216	0,800	T	304	0,200
rs41303343	CYP3A5(*7)	A	156	0,102	-	1366	0,898
rs55886062	DPYD(*13)	A	1528	1,000	C	0	**0,000**
rs67376798	DPYD	A	0	**0,000**	T	1524	1,000
rs3918290	DPYD(*2A)	C	1517	0,999	T	1	0,001

dbSNP	Gene (star allele)	VIC Allele	Count	Frequency	FAM Allele	Count	Frequency
rs115232898	*DPYD*	C	27	0,018	T	1499	0,982
rs2234922	*EPHX1*	A	1101	0,721	G	425	0,279
rs1051740	*EPHX1*	C	372	0,245	T	1148	0,755
rs3212986	*ERCC1*	A	357	0,234	C	1169	0,766
rs11615	*ERCC1*	A	78	0,051	G	1446	0,949
rs1047303	*HSD3B1*	C	161	0,106	A	1357	0,894
rs714368	*SLC22A16*	C	679	0,444	T	849	0,556
rs11045879	*SLCO1B1*	C	100	0,066	T	1426	0,934
rs4149117	*SLCO1B3*	G	433	0,285	T	1087	0,715
rs7311358	*SLCO1B3*	A	437	0,286	G	1089	0,714
rs11045585	*SLCO1B3*	A	1293	0,848	G	231	0,152
rs1042028	*SULT1A1(*2)*	C	1176	0,772	T	348	0,228
rs1142345	*TPMT*	C	67	0,044	T	1459	0,956
rs3828743	*TSPYL1*	A	62	0,041	G	1464	0,959
rs2070959	*UGT1A6*	A	1300	0,852	G	226	0,148
rs17863783	*UGT1A6*	G	1319	0,865	T	205	0,135
rs1105879	*UGT1A6*	A	1210	0,793	C	316	0,207
rs1902023	*UGT2B15*	A	477	0,312	C	1051	0,688
rs25487	*XRCC1*	C	1357	0,893	T	163	0,107

* Red numbers indicate the samples that show no genetic variation and alleles that are fixed within the population

Table 11. Application of the Hardy-Weinberg Exact Test to the 60 SNPs.

		Hardy-Weinberg Exact Test				
dbSNP	Gene (star allele)	Homozygous (VIC/VIC)	Heterozygous (VIC/FAM)	Homozygous (FAM/FAM)	D	*P*-value
rs2032582	*ABCB1*	CC (n) = 761	CT (n) = 0	TT (n) = 0	0	1
rs1045642	*ABCB1*	AA (n) = 9	AG (n) = 143	GG (n) = 611	-0.51	0.85
rs1128503	*ABCB1*	AA (n) = 5	AG (n) = 140	GG (n) = 618	2.37	0.42
rs3740066	*ABCC2*	CC (n) = 590	CT (n) = 159	TT (n) = 13	-1.77	0.5
rs2273697	*ABCC2*	AA (n) = 28	AG (n) = 197	GG (n) = 535	-6.94	0.09
rs717620	*ABCC2*	CC (n) = 740	CT (n) = 20	TT (n) = 0	0.13	1
rs225440	*ABCG1*	TT (n) = 276	TC (n) = 357	CC (n) = 127	-4.2	0.55
rs2231142	*ABCG2*	GG (n) = 756	GT (n) = 6	TT (n) = 0	0.01	1
rs9024	*CBR1*	AA (n) = 0	AG (n) = 7	GG (n) = 754	0.02	1
rs8133052	*CBR3*	AA (n) = 67	AG (n) = 306	GG (n) = 389	-3.48	0.54
rs1056892	*CBR3*	AA (n) = 172	AG (n) = 388	GG (n) = 201	4.03	0.61
rs2486758	*CYP17A1*	CC (n) = 1	CT (n) = 46	TT (n) = 715	-0.24	0.53
rs4646	*CYP19A1*	AA (n) = 126	AC (n) = 351	CC (n) = 285	-6.71	0.32
rs1048943	*CYP1A1*	CC (n) = 0	CT (n) =1	TT (n) = 762	0	1
rs2606345	*CYP1A1*	AA (n) = 1	AC (n) = 11	CC (n) = 749	-0.94	0.05
rs762551	*CYP1A2*	CC (n) 142	CA (n) = 378	AA (n) = 242	1.78	0.83
rs1056836	*CYP1B1*	CC (n) = 497	CG (n) 226	GG (n) = 37	-739	0.11
rs12721655	*CYP2B6(*8)*	AA (n) = 764	AG (n) = 0	GG (n) = 0	0	1
rs3211371	*CYP2B6(*5)*	CC (n) = 753	CT (n) = 8	TT (n) = 0	0.02	1
rs28399499	*CYP2B6(*16)*	CC (n) = 12	CT (n) = 152	TT (n) = 594	-1.78	0.48
rs3745274	*CYP2B6(*6)*	GG (n) = 311	GT (n) = 358	TT (n) = 93	4.09	0.58
rs4244285	*CYP2C19(*2)*	AA (n) = 25	AG (n) = 206	GG (n) = 530	-3.47	0.36
rs12248560	*CYP2C19(*17)*	CC (n) = 536	CT (n) = 207	TT (n) = 18	1.4	0.79
rs7902257	*CYP2C19(*27)*	AA (n) = 18	AG (n) = 193	GG (n) = 550	-0.77	0.78
rs10509681	*CYP2C8(*3)*	CC (n) = 0	CT (n) = 6	TT (n) = 754	0.01	1
rs1058930	*CYP2C8(*4)*	CC (n) = 0	CG (n) = 12	GG (n) = 749	0.05	1
rs11572103	*CYP2C8(*2)*	AA (n) = 31	AT (n) = 257	TT (n) = 476	2.3	0.66
rs7900194	*CYP2C9(*8)*	TT (n) = 0	TG (n) = 0	GG (n) = 757	0	1
rs1065852	*CYP2D6(*10)*	AA (n) = 28	AG (n) = 88	GG (n) = 632	-21.07	**<0.01**
rs3892097	*CYP2D6(*4)*	CC (n) = 713	CT (n) = 35	TT (n) = 5	-4.33	**<0.01**
rs35742686	*CYP2D6(*3)*	TT (n) = 749	T/- (n) = 0	-/- (n) = 1	-1	**<0.01**
rs28371706	*CYP2D6(*17)*	GG (n) = 477	GA (n) = 198	AA (n) = 70	-31.66	**<0.01**
rs4986907	*CYP3A4(*15)*	CC (n) = 730	CT (n) = 26	TT (n) = 1	-0.74	0.22
rs2740574	*CYP3A4(*1B)*	CC (n) = 432	CT (n) = 285	TT (n) = 46	0.57	1
rs35599367	*CYP3A4(*22)*	GG (n) = 761	GA (n) = 3	AA (n) = 0	0	1
rs776746	*CYP3A5(*3)*	TT (n) = 539	CT (n) = 202	CC (n) = 20	-0.76	0.79
rs10264272	*CYP3A5(*6)*	CC (n) = 488	CT (n) = 240	TT (n) = 32	-1.6	0.73
rs41303343	*CYP3A5(*7)*	AA (n) = 7	A/- (n) = 142	-/- (n) = 612	0.99	0.84
rs55886062	*DPYD(*13)*	AA (n) = 764	AC (n) = 0	CC (n) = 0	0	1
rs67376798	*DPYD*	AA (n) = 0	AT (n) = 0	TT (n) = 762	0	1
rs3918290	*DPYD(*2A)*	CC (n) = 758	CT (n) = 1	TT (n) = 0	0	1

dbSNP	Gene (star allele)	Homozygous (VIC/VIC)	Heterozygous (VIC/FAM)	Homozygous (FAM/FAM)	D	*P*-value
rs115232898	*DPYD*	CC (n) =1	CT (n) = 25	TT (n) = 737	-0.76	0.21
rs2234922	*EPHX1*	AA (n) = 398	AG (n) = 305	GG (n) = 60	-0.82	0.86
rs1051740	*EPHX1*	CC (n) = 42	CT (n) = 288	TT (n) = 430	3.52	0.56
rs3212986	*ERCC1*	AA (n) = 41	AC (n) = 275	CC (n) = 447	0.76	0.92
rs11615	*ERCC1*	AA (n) = 7	AG (n) = 64	GG (n) = 691	-5	**<0.01**
rs1047303	*HSD3B1*	CC (n) = 14	CA (n) = 133	AA (n) = 612	-5.46	0.05
rs714368	*SLC22A16*	CC (n) = 145	CT (n) = 389	TT (n) = 230	5.86	0.42
rs11045879	*SLCO1B1*	CC (n) = 5	CT (n) = 90	TT (n) = 668	-1.72	0.36
rs4149117	*SLCO1B3*	GG (n) = 58	GT (n) = 317	TT (n) = 385	3.67	0.59
rs7311358	*SLCO1B3*	AA (n) = 60	AG (n) = 317	GG (n) = 386	2.57	0.72
rs11045585	*SLCO1B3*	AA (n) = 550	AG (n) = 193	GG (n) = 19	-1.49	0.67
rs1042028	*SULT1A1(*2)*	CC (n) = 415	CT (n) = 346	TT (n) = 1	38.73	**<0.01**
rs1142345	*TPMT*	CC (n) = 2	CT (n) = 63	TT (n) = 698	-0.53	0.65
rs3828743	*TSPYL1*	AA (n) = 2	AG (n) = 58	GG (n) = 703	-0.74	0.36
rs2070959	*UGT1A6*	AA (n) = 551	AG (n) = 198	GG (n) = 14	2.74	0.57
rs17863783	*UGT1A6*	GG (n) = 573	GT (n) = 173	TT (n) = 16	-2.21	0.53
rs1105879	*UGT1A6*	AA (n) = 476	AC (n) = 258	CC (n) = 29	3.72	0.51
rs1902023	*UGT2B15*	AA (n) = 84	AC (n) = 309	CC (n) = 371	-9.55	0.11
rs25487	*XRCC1*	CC (n) = 607	CT (n) = 143	TT (n) = 10	-1.26	0.57

* The interim statistic used to calculate the *P*-value. The larger the positive or negative this value is, the smaller or more significant the *P*-value. Significant P-values are indicated in red.

Table 12. NCBI SNP database flanking sequences.[102]

SNP ID	Flanking Sequence
CYP2D6*10 (rs1065852)	TCCAGGACCTCCTCCCTCACCTGGTCGAAGCAGTATGGTGTGTTCTGGAA GTCCACATGCAGCAGGTTGCCCAGCCCGGGCAGTGGCAGGGGGCCTGGTG [G/A/C] GTAGCGTGCAGCCCAGCGTTGGCGCCGGTGCATCAGGTCCACCAGGAGCA GGAAGATGGCCACTATCACGGCCAGGGGCACCAGTGCTTCTAGCCCCATA
CYP2D6*4 (rs3892097)	AAGCGAGGGTCGTCGTACTCGAAGCGGCGCCCGCAGGTGAGGGAGGCGAT CACGTTGCTCACGGCTTTGTCCAAGAGACCGTTGGGGCGAAAGGGGCGTC [C/T] TGGGGGTGGGAGATGCGGGTAAGGGGTCGCCTTCCCCGTCCCCCGCCTTC CCAGTTCCCGCTTTGTGCCCTTCTGCCCATCACCCACCGGAGTGGTTGGC
CYP2D6*3 (rs35742686)	CCCTTGCCCCCCACCGTGGCAGCCACTCTCACCTTCTCCATCTCTGCCAG GAAGGCCTCAGTCAGGTCTCGGGGGGGCTGGGCTGGGTCCCAGGTCATCC [T/-] GTGCTCAGTTAGCAGCTCATCCAGCTGGGTCAGGAAAGCCTTTTGGAAGC GTAGGACCTTGCCAGCCAGCGCTGGGATATGCAGGAGGACGGGGACAGCA
CYP2D6*17 (rs28371706)	AGCTCGGACTACGGTCATCACCCACCCGGGTCCCACGGAAATCTGTCTCT GTCCCCACCGCTGCTTGCCTTGGGAACGCGGCCCGAAACCCAGGATCTGG [G/A/C/T] TGATGGGCACAGGCGGGCGGTCGGCGGTGTCCTCGCCGTGGGTCACCAGC GCCTCGCGCACGGCCGCCAGCCCATTGAGCACGACCACCGGCGTCCAGGC
ERCC1 (rs11615)	GAAGTCTGGGGTGGCGCCGCAGAGCTCACCTGAGGAACAGGGCACAGGTG CTCTGGCCCAGCACATAGTCGGGAATTACGTCGCCAAATTCCCAGGGCAC [A/G] TTGCGCACGAACTTCAGTACGGGATTGCCCCTCTGGGGAGGGACGAAGGG CAGAAGCCATCAATAGGGATGACCCTTGATAACCACAGGGCCCTCCTCCA
SULT1A1 (rs1042028)	CTCCTGGGGGACGGTGGTGTAGTTGGTCATAGGGTTCTTCTTCATCTCCT TGAACGACGTGTGCTGAACCACGAAGTCCACGGTCTCCTCTGGCAGGGAG [C/T] GCCCCACAAACTCCAGGATCTTTTGAATCTCCCTTTTCGGGTTCTGAGCA GCAGAGGGCCCCTCAGTGGAGGCTCGGATTACTGATTCAGGAAAAGTAAA

The four *CYP2D6* SNPs flanking sequences aligned to several other gene regions with 100% coverage. The distance trees of the query sequence and the database sequences were generated using the BLAST pairwise alignment. The tree reflects the alignment of the sequencing where 100% homology is attained, no distance between the nodes was observed. This principle was applied to all the SNPs showing Hardy-Weinberg deviation (HWD).[118] The high degree of homology reduces the specificity of the genotyping assay for *CYP2D6*, potentially leading to unreliable genotyping classification, and clinical association.

The *CYP2D6* gene is highly polymorphic with over 100 recognised star alleles. The *CYP2D6* locus has a high degree of variation and homology to pseudogenes. A high sequencing identify implies that it is highly likely that the two aligned sequences diverged from a common ancestral sequence (and are thus homologous), and that it did not evolve independently.[122] Given the complexity of the *CYP2D6* gene, along with the high degree of genetic diversity owing to duplications, deletions, tandem repeat and copy number changes, deviation from HW can be explained by the non-specificity of probes (Figures 39 to 42, Appendix VII). No further analysis was performed on the *CYP2D6* gene. For this gene region, targeted gene sequencing or whole exome sequencing would be essential, due to the high degree of variation and detection of potential novel variants. Furthermore, the study focused on the prevalence of pharmacogenetic markers and was not intended to correlate it to real treatment outcomes, therefore the need to do any investigation into copy number variation was not required.

The remaining two SNPs, rs11615 (*ERCC1*) and rs1042028 (*SULT1A1*), also showed 100% homology to other regions of the genome, however to a far lesser extent. The probe for rs1042028 was custom designed and had not been previously tested nor validated by the manufacturers. The close clustering of the samples between the heterozygotes and homozygotes for the VIC allele could also have resulted in genotyping errors, even though the analysis was double checked.

Population stratification could also cause the absence of HWE, and this was further investigated by removing certain cohorts from the study and repeating the Hardy-Weinberg test. *SULT1A1*2*, has been associated with an increased risk of BrCa in Asian women.[123] *ERCC1* (rs11615) has also been linked to increased susceptibility to BrCa. The DNA repair protein is vital for maintaining genomic fidelity and integrity.[124] Sampling bias, is a bias that is introduced when sample collection is done in such a manner that the members of the intended population have a lower or higher sampling probability. Sampling bias could also not be excluded, even with the removal of certain populations there was still discordance with HW. The study consisted of four cohorts, from various region within South Africa, reducing the introduction of sample bias.

Taking the above into account, various cohorts were removed, and the HWE test repeated. Firstly, the PrCa cohort was removed from the analysis of rs11615 and rs1042028 (P = 0.0025 and P < 0.0001, respectively), and thereafter the BrCa cohort (P = 0.0007 and P < 0.0001, respectively). Despite this, a significant discrepancy in terms of HWE was still noted. To confirm the finding, both the BrCa and PrCa cohorts were removed (P = 0.000235 and P = 0.00, respectively). The results are indicative thereof that population stratification is an unlikely contributor to the HWD.

To investigate the probability of population stratification, the two cancer study cohorts were removed and the HWE test was repeated. However, even with the two cohorts removed, the four *CYP2D6* SNPs still did not conform to the HWE, kindly refer to Appendix VIII calculations.

4.6 Allele frequency comparison between the study population and previously reported allele frequencies

The 1000 Genomes Project was an international research effort among various research groups worldwide. The population groups included were i) the European population (EUR) which consisted of people from Toscani in Italy, Finland, from the United Kingdom (England and Scotland) as well as Spain, ii) the African (AFR) population which included people from Yoruba and Esam in Nigeria, Wuhya in Kenya, Gambia, and Mende in Sierre Leone, as well as iii) the Americans of African Ancestor (USA) and African Caribbean's (Barbados) (Figure 25). Little to no definite genotypic data is available for SSA, which includes South Africa.[110] What the study did highlight was the diversity of the ADME landscape, and the importance of further intra-population based studies.

The 1000 Genomes Project also aimed to provide a resource of almost all variants, including SNPs and structural variants as well as their haplotype contexts for the population groups included, it remains the most comprehensive database.[109] The product was a resource that enabled genome-wide association studies to focus on almost all variants that exist in regions found to be associated with disease. A Chi-square test in conjunction with the Fisher Exact test, was conducted to highlight the deviation in the observed allele frequency in the study population with population data

previously reported in the 1000 Genomes database for population of African and European ancestry.
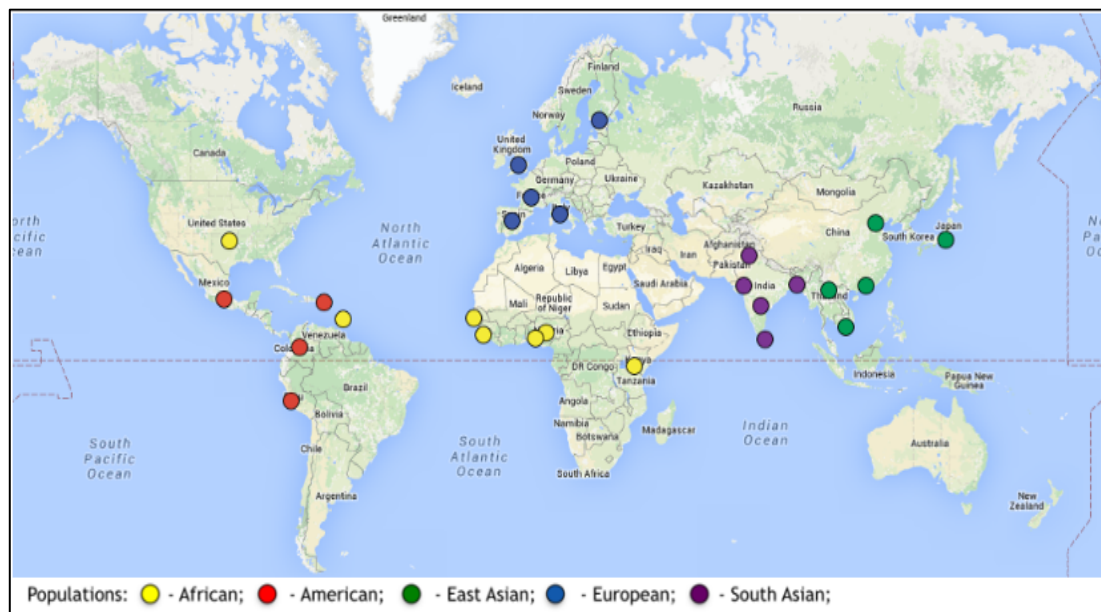

Figure 25. Geographical representation of the populations included in the 1000 Genomes Project.[110]

Population size is important when conducting population-based studies. The larger the population size, the greater the statistical power of the analysis. Two tests for independence were employed to determine the level of significance between the study population and population of African and European ancestry as per the allele frequencies reported in the 1000 Genomes Project. The 1000 Genomes Project database remains one of the largest and most comprehensive, including over 2000 samples from 26 population groups. The degrees of freedom (DF) for the Chi-square test were 1. The Chi-square and Fisher Exact test are both tests of independence; the population size plays an important factor in each test. Tables 13 and 14 indicate the level of significance between the populations at a 95% confidence interval ($P < 0.05$). Both statistical tests resulted in the same or very similar P-values, confirming the level of significance.

A total of 29 SNPs showed a significant ($P < 0.05$) difference between the observed allele frequencies and those previously reported for the population of African ancestry. This suggests that the self-identified black South African study population (current

study) is significantly different in terms of allele frequency compared to data collected during the 1000 Genomes Project for 48% of the SNPs investigated (Table 13). Due to these differences, there is clearly a need for further investigation in the SSA population as well as intra-population studies.

When comparing the allele frequencies of the current study group to the European population from the 1000 Genomes Project it was found that 53/60 SNPs (88.3%) showed a significant ($P < 0.05$) deviation (Table 14). Most of the drugs on the market today, have been tested in the European or Asian population, with limited information being available regarding the effect of drugs on the African population and more so the SSA population, which has shown to be significantly different in terms of allele frequencies.[94] It is evident that drug trials need to include a wider diversity of populations. Building onto this, is the need for improved pharmacovigilance training and reporting, so that side effects that may be population specific can also be investigated. This is essential to close the information gap, so that reported cases can be investigated where necessary.

Table 13. Comparison of current study population and the African population in the 1000 Genomes Project.

dbSNP	Gene (star allele)	VIC Allele	n	FAM Allele	n	VIC Allele (n)	FAM Allele (n)	x^2	P-value	P-value
				Observed Allele		1000 Genomes AFR		Chi-square Test (DF = 1)		Fisher Exact Test
rs2032582	*ABCB1*	C	1540	T	0	1295	1	0.01	= 0.09	= 1
rs1045642	*ABCB1*	A	161	G	1365	198	1124	12.57	**< 0.01**	**< 0.01**
rs1128503	*ABCB1*	A	150	G	1376	180	1142	9.88	**< 0.01**	**< 0.01**
rs3740066	*ABCC2*	C	1339	T	185	1036	286	46.24	**< 0.01**	**< 0.01**
rs2273697	*ABCC2*	A	253	G	1267	250	1072	2.49	= 0.11	= 0.12
rs717620	*ABCC2*	C	1500	T	20	1281	41	10.84	**< 0.01**	**< 0.01**
rs225440	*ABCG1*	T	909	C	611	784	538	0.07	= 0.79	= 0.79
rs2231142	*ABCG2*	G	1518	T	6	1305	17	7.22	**< 0.01**	**= 0.01**
rs9024	*CBR1*	A	7	G	1515	9	1314	0.61	= 0.43	= 0.46
rs8133052	*CBR3*	A	440	G	1084	324	998	6.88	**< 0.01**	**< 0.01**
rs1056892	*CBR3*	A	732	G	790	674	648	2.36	= 0.12	= 0.13
rs2486758	*CYP17A1*	C	48	T	1476	59	1263	3.36	= 0.07	= 0.08
rs4646	*CYP19A1*	A	603	C	921	376	946	39.1	**< 0.01**	**< 0.01**
rs1048943	*CYP1A1*	C	1	T	1525	9	1313	8.59	**< 0.01**	**< 0.01**
rs2606345	*CYP1A1*	A	13	C	1509	66	1256	47.97	**< 0.01**	**< 0.01**
rs762551	*CYP1A2*	C	662	A	862	579	743	0.04	= 0.85	= 0.85
rs1056836	*CYP1B1*	C	1220	G	300	1080	242	0.94	= 0.33	= 0.34
rs12721655	*CYP2B6(*8)*	A	1528	G	0	1320	2	3.07	= 0.08	= 0.022
rs3211371	*CYP2B6(*5)*	C	1514	T	8	1307	15	3.29	= 0.07	= 0.09
rs28399499	*CYP2B6(*16)*	C	176	T	1340	109	1213	8.95	**< 0.01**	**< 0.01**
rs3745274	*CYP2B6(*6)*	G	980	T	544	827	495	0.93	= 0.33	= 0.35
rs4244285	*CYP2C19(*2)*	A	256	G	1266	225	1097	0.02	= 0.89	= 0.92
rs12248560	*CYP2C19(*17)*	C	1279	T	243	1011	311	25.73	**< 0.01**	**< 0.01**
rs7902257	*CYP2C19(*27)*	A	229	G	1293	109	1213	32.02	**< 0.01**	**< 0.01**

dbSNP	Gene (star allele)	VIC Allele	n	FAM Allele	n	1000 Genomes AFR VIC Allele (n)	FAM Allele (n)	Chi-square Test (DF = 1) x^2	P-value	Fisher Exact Test P-value
rs10509681	CYP2C8(*3)	C	6	T	1514	11	1311	2.29	= 0.13	= 0.15
rs1058930	CYP2C8(*4)	C	12	G	1510	5	1317	2.08	= 0.15	= 0.22
rs11572103	CYP2C8(*2)	A	319	T	1209	250	1072	1.72	= 0.19	= 0.20
rs7900194	CYP2C9(*8)	T	0	G	1514	70	1252	108.89	**< 0.01**	**< 0.01**
rs1065852	CYP2D6(*10)	A	144	G	1352	149	1173	2.03	= 0.15	= 0.16
rs3892097	CYP2D6(*4)	C	1461	T	45	1242	80	15.72	**< 0.01**	**< 0.01**
rs35742686	CYP2D6(*3)	T	1498	-	2	1319	3	0.35	= 0.56	= 0.67
rs28371706	CYP2D6(*17)	G	1152	A	338	1034	288	0.33	= 0.57	= 0.59
rs4986907	CYP3A4(*15)	C	1486	T	28	1299	23	0.05	= 0.83	= 0.89
rs2740574	CYP3A4(*1B)	C	1149	T	377	1012	310	0.61	= 0.43	= 0.46
rs35599367	CYP3A4(*22)	G	1525	A	3	1321	1	0.78	= 0.38	= 0.63
rs776746	CYP3A5(*3)	T	1280	C	242	1084	238	2.23	= 0.14	= 0.15
rs10264272	CYP3A5(*6)	C	1216	T	304	1118	204	10.12	**< 0.01**	**< 0.01**
rs41303343	CYP3A5(*7)	A	156	-	1366	156	1166	1.74	= 0.19	= 0.21
rs55886062	DPYD(*13)	A	1528	C	0	1322	0	0	= 1.00	= 1.00
rs67376798	DPYD	A	0	T	1524	1	1321	1.53	= 0.22	= 0.46
rs3918290	DPYD(*2A)	C	1517	T	1	1321	1	0.01	= 0.92	= 1.00
rs115232898	DPYD	C	27	T	1499	30	1292	0.9	= 0.34	= 0.35
rs2234922	EPHX1	A	1101	G	425	855	467	18.37	**< 0.01**	**< 0.01**
rs1051740	EPHX1	C	372	T	1148	186	1136	49.46	**< 0.01**	**< 0.01**
rs3212986	ERCC1	A	357	C	1169	385	937	12.04	**< 0.01**	**< 0.01**
rs11615	ERCC1	A	78	G	1146	47	1273	10.82	**< 0.01**	**< 0.01**
rs1047303	HSD3B1	C	161	A	1357	113	1209	3.46	= 0.06	= 0.06
rs714368	SLC22A16	C	679	T	849	514	808	9.01	**< 0.01**	**< 0.01**
rs11045879	SLCO1B1	C	100	T	1426	250	1075	101.66	**< 0.01**	**< 0.01**
rs4149117	SLCO1B3	G	433	T	1087	471	851	16.6	**< 0.01**	**< 0.01**

dbSNP	Gene (star allele)	VIC Allele	n	FAM Allele	n	1000 Genomes AFR VIC Allele (n)	FAM Allele (n)	Chi-square Test (DF = 1) x^2	P-value	Fisher Exact Test P-value
rs7311358	SLCO1B3	A	437	G	1089	470	852	15.59	< 0.01	< 0.01
rs11045585	SLCO1B3	A	1293	G	231	1026	296	24.5	< 0.01	< 0.01
rs1042028	SULT1A1(*2)	C	1176	T	348	2585	873	3.35	= 0.07	= 0.07
rs1142345	TPMT	C	67	T	1459	88	1234	7.05	< 0.01	< 0.01
rs3828743	TSPYL1	A	62	G	1464	85	1237	8.09	< 0.01	< 0.01
rs2070959	UGT1A6(*2)	A	1300	G	226	1006	316	37.97	< 0.01	< 0.01
rs17863783	UGT1A6(*4)	G	1319	T	205	1151	171	0.16	= 0.68	= 0.70
rs1105879	UGT1A6(*3)	A	1210	C	316	923	399	33.77	< 0.01	< 0.01
rs1902023	UGT2B15(*2)	A	477	C	1051	526	796	22.81	< 0.01	< 0.01
rs25487	XRCC1	C	1357	T	163	1176	146	0.07	= 0.78	= 0.81

* Red numbers reflect significant P-values

Table 14. Comparison of current study population and the European population in the 1000 Genomes Project.

dbSNP	Gene (star allele)	VIC Allele	n	FAM Allele	n	1000 Genomes EUR VIC Allele (n)	FAM Allele (n)	Chi-square Test (DF = 1) χ^2	P-value	Fisher Exact Test P-value
rs2032582	*ABCB1*	C	1540	T	0	573	18	39.24	**< 0.01**	**< 0.01**
rs1045642	*ABCB1*	A	161	G	1365	512	485	516.69	**< 0.01**	**< 0.01**
rs1128503	*ABCB1*	A	150	G	1376	418	588	349.3	**< 0.01**	**< 0.01**
rs3740066	*ABCC2*	C	1339	T	185	633	373	216.24	**< 0.01**	**< 0.01**
rs2273697	*ABCC2*	A	253	G	1267	205	801	5.63	= 0.02	= 0.02
rs717620	*ABCC2*	C	1500	T	20	798	208	293.13	**< 0.01**	**< 0.01**
rs225440	*ABCG1*	T	909	C	611	412	594	86.58	**< 0.01**	**< 0.01**
rs2231142	*ABCG2*	G	1518	T	6	911	95	141	**< 0.01**	**< 0.01**
rs9024	*CBR1*	A	7	G	1515	88	918	123.64	**< 0.01**	**< 0.01**
rs8133052	*CBR3*	A	440	G	1084	455	551	70.41	**< 0.01**	**< 0.01**
rs1056892	*CBR3*	A	732	G	790	356	650	40.23	**< 0.01**	**< 0.01**
rs2486758	*CYP17A1*	C	48	T	1476	227	779	238.76	**< 0.01**	**< 0.01**
rs4646	*CYP19A1*	A	603	C	921	292	714	29.82	**< 0.01**	**< 0.01**
rs1048943	*CYP1A1*	C	1	T	1525	35	971	57.2	**< 0.01**	**< 0.01**
rs2606345	*CYP1A1*	A	13	C	1509	667	339	1508.44	**< 0.01**	**< 0.01**
rs762551	*CYP1A2*	C	662	A	862	322	684	33.67	**< 0.01**	**< 0.01**
rs1056836	*CYP1B1*	C	1220	G	300	400	606	434.99	**< 0.01**	**< 0.01**
rs12721655	*CYP2B6(*8)*	A	1528	G	0	1005	1	1.85	= 0.17	= 0.40
rs3211371	*CYP2B6(*5)*	C	1514	T	8	893	113	164.8	**< 0.01**	**< 0.01**
rs28399499	*CYP2B6(*16)*	C	176	T	1340	0	1006	187.85	**< 0.01**	**< 0.01**
rs3745274	*CYP2B6(*6)*	G	980	T	544	769	237	42.7	**< 0.01**	**< 0.01**
rs4244285	*CYP2C19(*2)*	A	256	G	1266	146	860	2.43	= 0.12	= 0.13

dbSNP	Gene (star allele)	VIC Allele	n	FAM Allele	n	1000Genome EUR VIC Allele (n)	FAM Allele (n)	Chi-square Test (DF = 1) χ^2	P-value	Fisher Exact Test P-value
rs12248560	CYP2C19(*17)	C	1279	T	243	781	225	16.22	< 0.01	< 0.01
rs7902257	CYP2C19(*27)	A	229	G	1293	1	1005	236.1	< 0.01	< 0.01
rs10509681	CYP2C8(*3)	C	6	T	1514	119	887	185.47	< 0.01	< 0.01
rs1058930	CYP2C8(*4)	C	12	G	1510	58	948	56.47	< 0.01	< 0.01
rs11572103	CYP2C8(*2)	A	319	T	1209	4	1002	315.8	< 0.01	< 0.01
rs7900194	CYP2C9(*8)	T	0	G	1514	2	1004	3.68	= 0.06	= 0.16
rs1065852	CYP2D6(*10)	A	144	G	1352	203	803	54.91	< 0.01	< 0.01
rs3892097	CYP2D6(*4)	C	1461	T	45	819	187	176.41	< 0.01	< 0.01
rs35742686	CYP2D6(*3)	T	1498	-	2	987	19	23.71	< 0.01	< 0.01
rs28371706	CYP2D6(*17)	G	1152	A	338	1004	2	362.53	< 0.01	< 0.01
rs4986907	CYP3A4(*15)	C	1486	T	28	1005	1	21.86	< 0.01	< 0.01
rs2740574	CYP3A4(*1B)	C	1149	T	377	28	978	1535.51	< 0.01	< 0.01
rs35599367	CYP3A4(*22)	G	1525	A	3	956	50	73.78	< 0.01	< 0.01
rs776746	CYP3A5(*3)	T	1280	C	242	57	949	1724.85	< 0.01	< 0.01
rs10264272	CYP3A5(*6)	C	1216	T	304	1003	3	307	< 0.01	< 0.01
rs41303343	CYP3A5(*7)	A	156	-	1366	0	1006	165.04	< 0.01	< 0.01
rs55886062	DPYD(*13)	A	1528	C	0	1005	1	1.85	= 0.17	= 0.40
rs67376798	DPYD	A	0	T	1524	7	999	12.94	< 0.01	< 0.01
rs3918290	DPYD(*2A)	C	1517	T	1	1001	5	4.82	= 0.03	= 0.04
rs115232898	DPYD	C	27	T	1499	0	1006	27.54	< 0.01	< 0.01
rs2234922	EPHX1	A	1101	G	425	841	165	45.96	< 0.01	< 0.01
rs1051740	EPHX1	C	372	T	1148	306	700	10.81	< 0.01	< 0.01
rs3212986	ERCC1	A	357	C	1169	252	754	0.91	= 0.34	= 0.34
rs11615	ERCC1	A	78	G	1146	626	380	866.92	< 0.01	< 0.01
rs1047303	HSD3B1	C	161	A	1357	342	664	204.55	< 0.01	< 0.01
rs714368	SLC22A16	C	679	T	849	222	784	137.21	< 0.01	< 0.01

86

dbSNP	Gene (star allele)	VIC Allele	n	FAM Allele	n	1000 Genomes EUR		Chi-square Test (DF = 1)		Fisher Exact Test
						VIC Allele (n)	FAM Allele (n)	χ^2	*P*-value	*P*-value
rs11045879	*SLCO1B1*	C	100	T	1426	191	815	90.1	**< 0.01**	**< 0.01**
rs4149117	*SLCO1B3*	G	433	T	1087	870	136	885.85	**< 0.01**	**< 0.01**
rs7311358	*SLCO1B3*	A	437	G	1089	870	136	882.67	**< 0.01**	**< 0.01**
rs11045585	*SLCO1B3*	A	1293	G	231	859	147	0.14	= 0.71	= 0.73
rs1042028	*SULT1A1(*2)*	C	1176	T	348	12269	6197	77.95	**< 0.01**	**< 0.01**
rs1142345	*TPMT*	C	67	T	1459	29	977	3.91	= 0.05	= 0.06
rs3828743	*TSPYL1*	A	62	G	1464	263	743	266.14	**< 0.01**	**< 0.01**
rs2070959	*UGT1A6(*2)*	A	1300	G	226	694	312	93.37	**< 0.01**	**< 0.01**
rs17863783	*UGT1A6(*4)*	G	1319	T	205	983	23	109.37	**< 0.01**	**< 0.01**
rs1105879	*UGT1A6(*3)*	A	1210	C	316	671	335	49.63	**< 0.01**	**< 0.01**
rs1902023	*UGT2B15(*2)*	A	477	C	1051	516	490	102.23	**< 0.01**	**< 0.01**
rs25487	*XRCC1*	C	1357	T	163	638	368	240.96	**< 0.01**	**< 0.01**

Red numbers reflect significant *P*-values

The CA makes use of the difference between the distribution of the groups, which is measured by the X^2 – distances. This is graphically represented in Figures 26 to 30. In these Figures, the alleles are indicated by the blue circles (row coordinates) on the left of the plot, and the population and study groups by the red squares on the right of the plot (column coordinates). The relation of the red squares in proportion to the blue circles, provides an indication of allele frequency and relatedness of the population groups.

CA plots for five SNPs could not be generated due to the lack of variance, these were rs2032582 (ABCB1), rs12721655 (*CYP2B6*8*), rs55886062 (*DPYD*13)*, rs67376798 (*DPYD*) and rs3918290 (*DPYD*2A*). Six SNPs showed minor variation between all three of the groups, as the X^2 – distances were very small, rs2273697 (*ABCC2)* and rs4244285 (*CYP2C19*2*) (Figure 32). The *DPYD* genes also showed little variation between the three groups, suggesting that drugs that are metabolised by these genes are metabolised in a similar manner. Only the AA genotype was observed for rs55886062, and TT for rs6737698 in the current study population. The AT genotype was noted in only one African group, but in seven observations in the European group. The major genotype for rs3918290 was CC, with 1 CT genotype being reported in this study group as well as the African group and five in the European group. A similar observation was noted for rs12721655, where only the AA genotype was recorded. The AG genotyped was observed twice in the African group and once in the European group.

Two SNPs showed closer clustering of the study group with the European group, this is represented by the overlapping red squares of the two groups: rs7900194 (*CYP2C9*8*), and rs11045585 (*SLCO1B3*) (Figure 27). For the study group, only the genotype GG was noted for rs7900194, whereas the African group had heterozygotes (GT) at the highest frequency, thus its position was closer to the "T". No homozygotes for the T allele were found for any of the groups. For rs11045585, the study group and European group clustered together as the allele frequency for the A was 0.848 and 0.854, respectively, while the allele frequency for the African group was 0.776. It is evident that the African group had the highest allele frequency for the G allele at 0.224 and is thus positioned closer to the G allele on the 1D plot.

Three SNPs showed closer clustering between the African and European group, with the study group being a significant distance away: rs4646 (*CYP19A1*), rs12248560 (*CYP2C19*17*) and rs11045879 (*SLCO1B1*) (Figure 28). The African and European populations from the 1000 Genomes Project would show similar drug efficacy, when taking drugs metabolised by the above three genes. However, the study cohort is likely to show varied efficacy, in comparison when prescribed the same drug.

Ten of the 60 SNPs showed little to no relationship between all three groups, with the CA plot reflecting a distance between all three the data points: rs374006 (*ABCC2*), rs28399499 (*CYP2B6*16*), rs7902557 (*CYP2C19*27*), rs2234922 (*EPHX1*), rs1051740 (*EPHX1*), rs714368 (*SLC22A1*), rs1142345 (*TPMT*), rs2070959 (*UGT1A6*2*), rs1105879 (*UGT1A6*3*), rs1902023 (*UGT2B15*2*). Figure 35 shows the CA plot for rs374006 (*ABCC2*) and rs28399499 (*CYP2B6*16*) as an example.

The remaining SNPs (35/60) showed closer clustering of the study group with the African group, and a distance relationship to the European group. The overlap of the red dots is indicative of a lack of significance between the two groups. However, in some instances the two red squares were in close proximity, which is still indicative of variation, but to a lesser extent than above. The CA plot for two examples; rs1045642 (*ABCB1*) and rs9024 (*CBR1*) is provided in Figure 30.

To graphically illustrate the relationship between the study group as well as the comparative groups from the 1000 Genomes Project (African and European), a principal component analysis (PCA) was performed on the minor allele frequencies of the three populations (Figure 37). The PCA determines the best-fit line, which is the average squared distance from the points in the line. A best-fit line was drawn for each population that was compared. The distance between the lines is an indication of the variability between the populations. The PCA plot reiterates what the CA plots illustrated, expect the CA is a descriptive technique, that is applied to tables of which the chi-squared statistic is appropriate. The analysis showed difference in distribution between the populations, however, the study and African population clustered closer together in comparison to the European population. The plot also highlighted the SNPs which are more prominent in the various population.
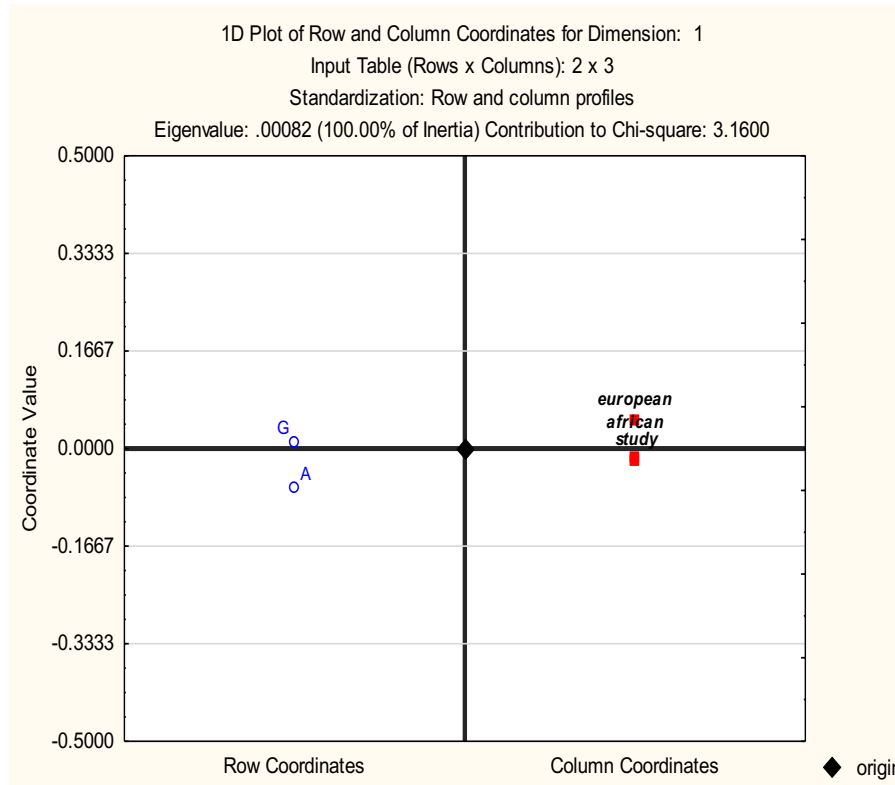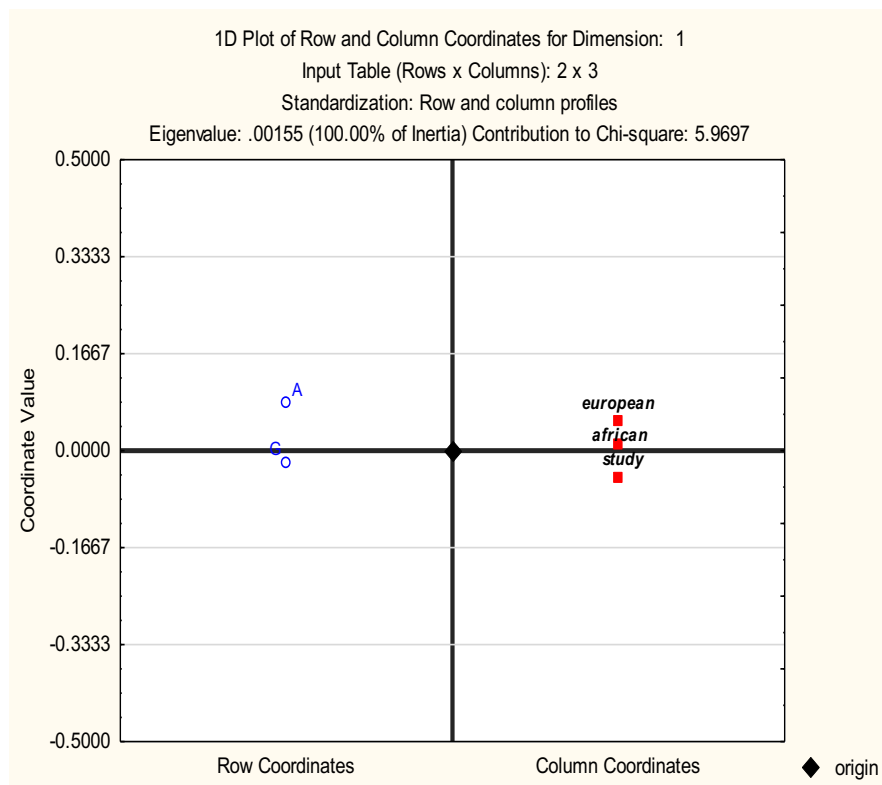
Figure 26. Correspondence analysis (CA) plot of rs2273697 (*ABCC2*) left and rs4244285 (*CYP2C19*2*) right. The distance between the points were not significant, suggesting the populations are very similar in allele frequencies. The blue circles represent the allele or row coordinates, while the red squares represent the three populations being compared.
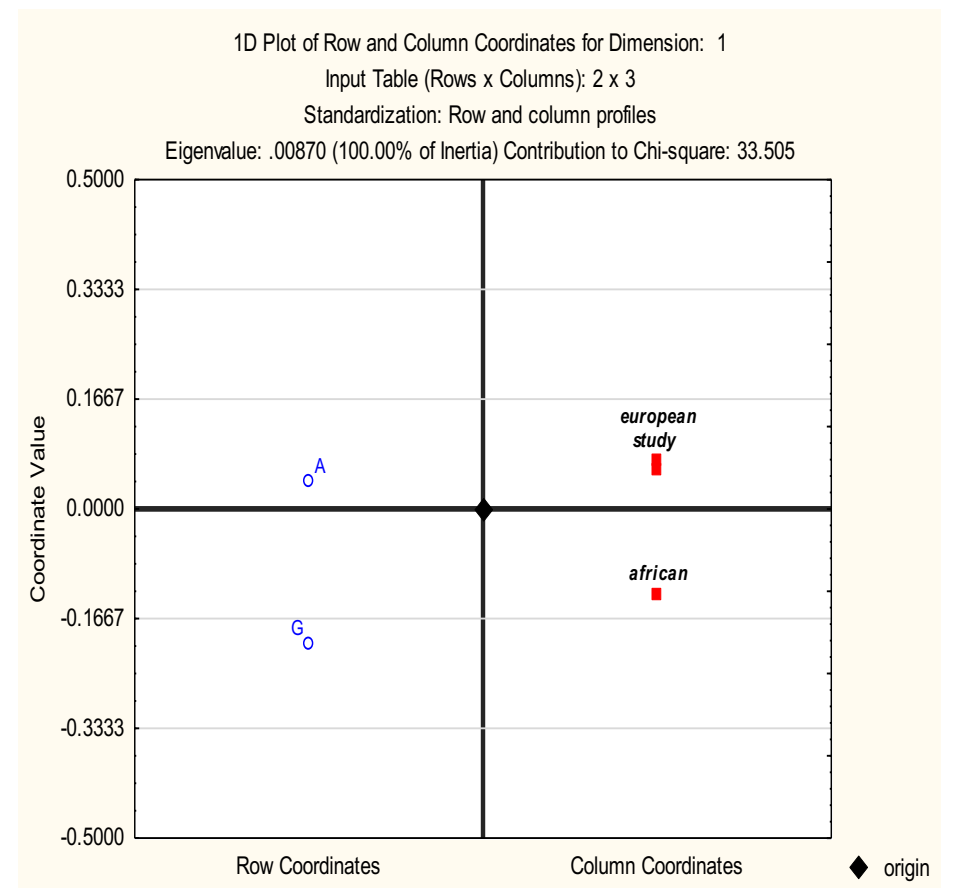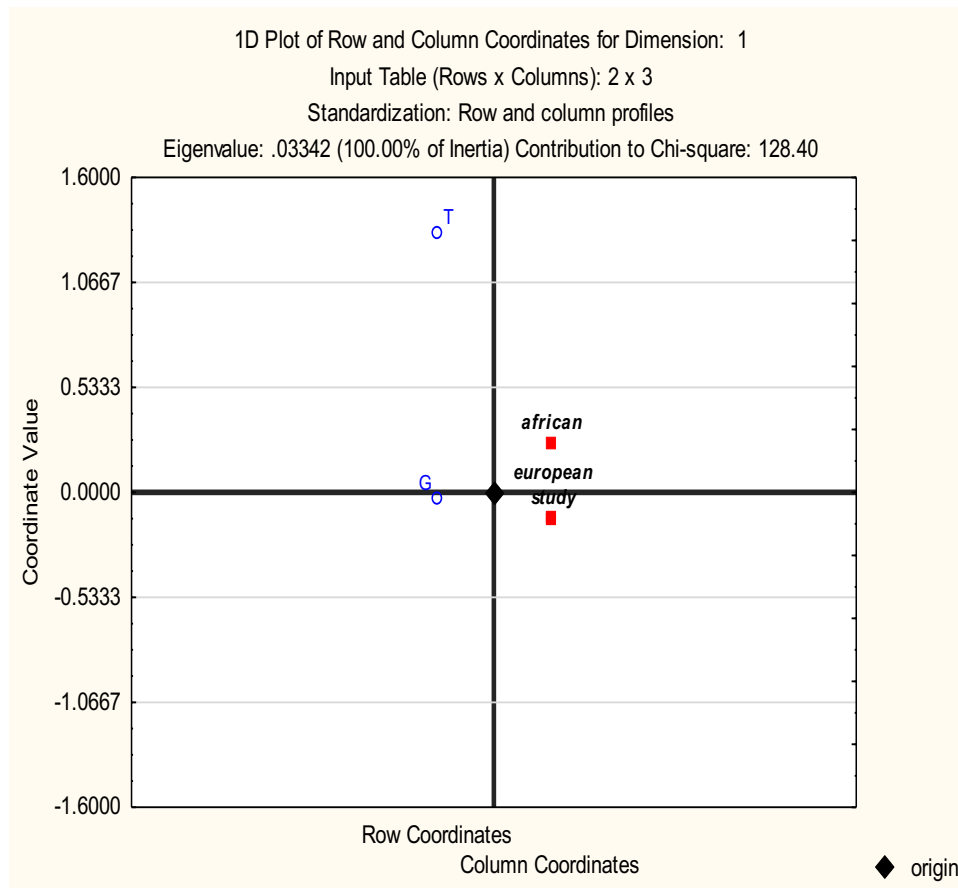
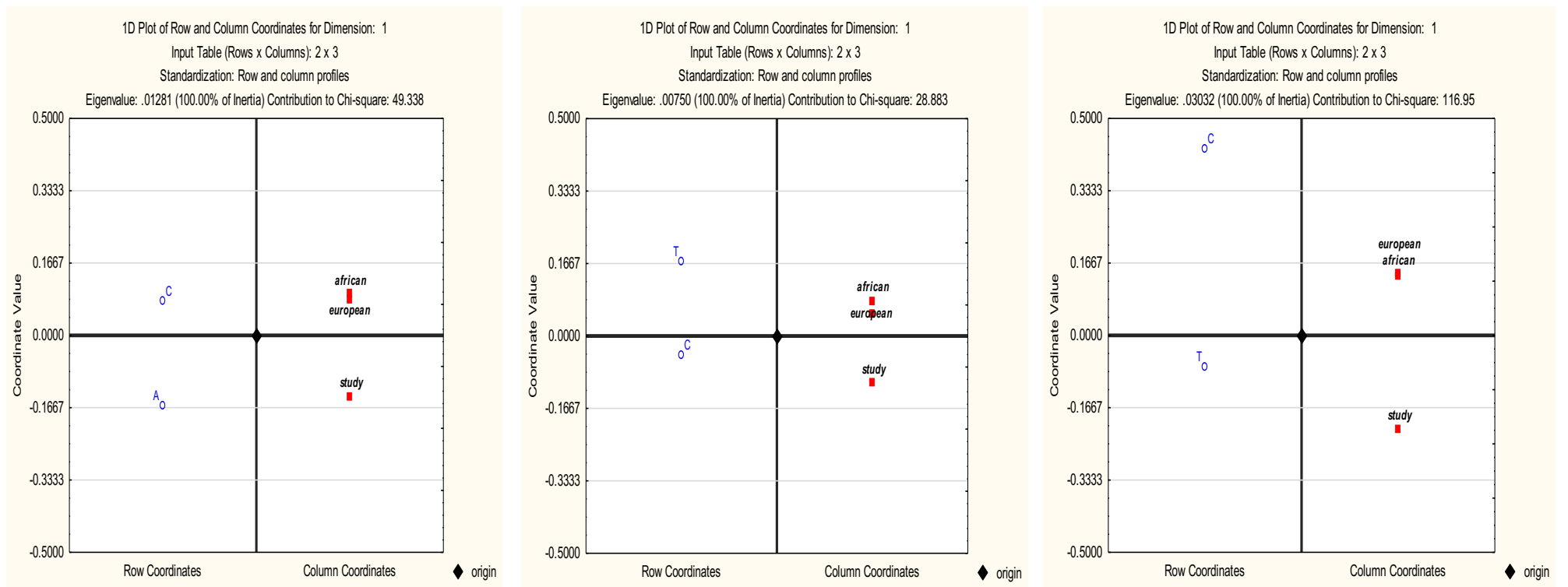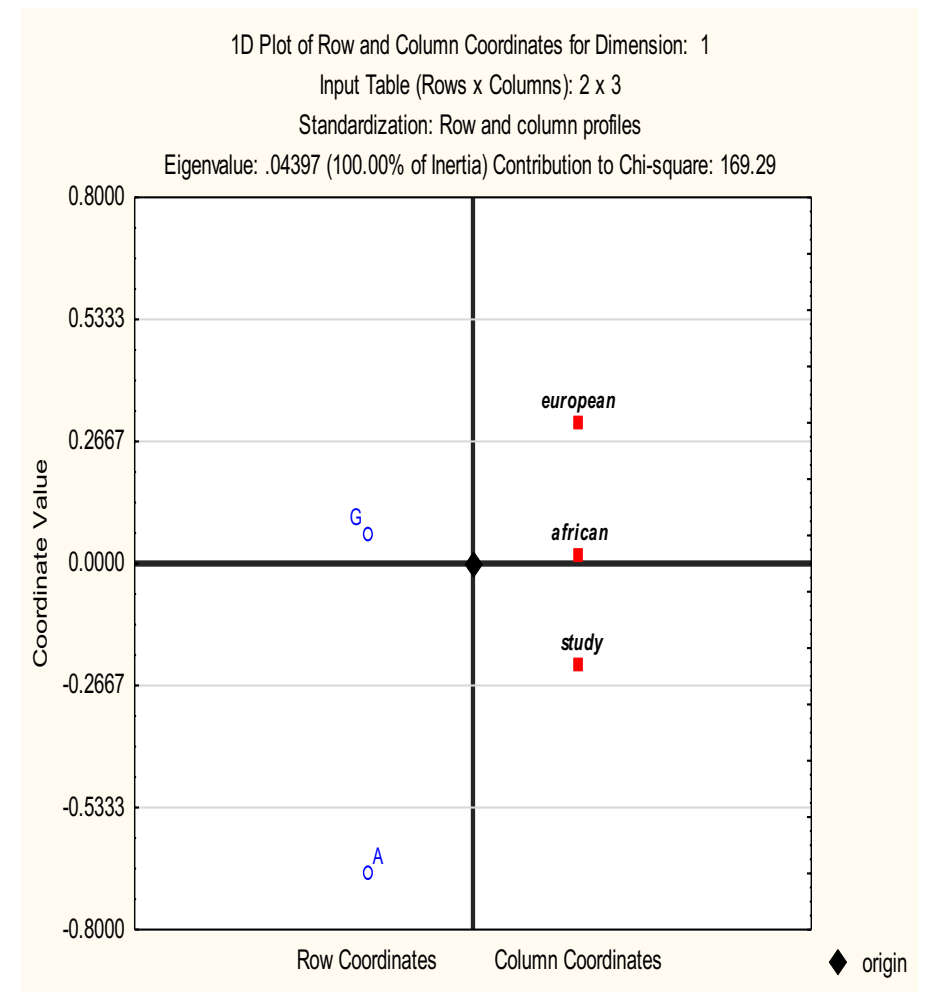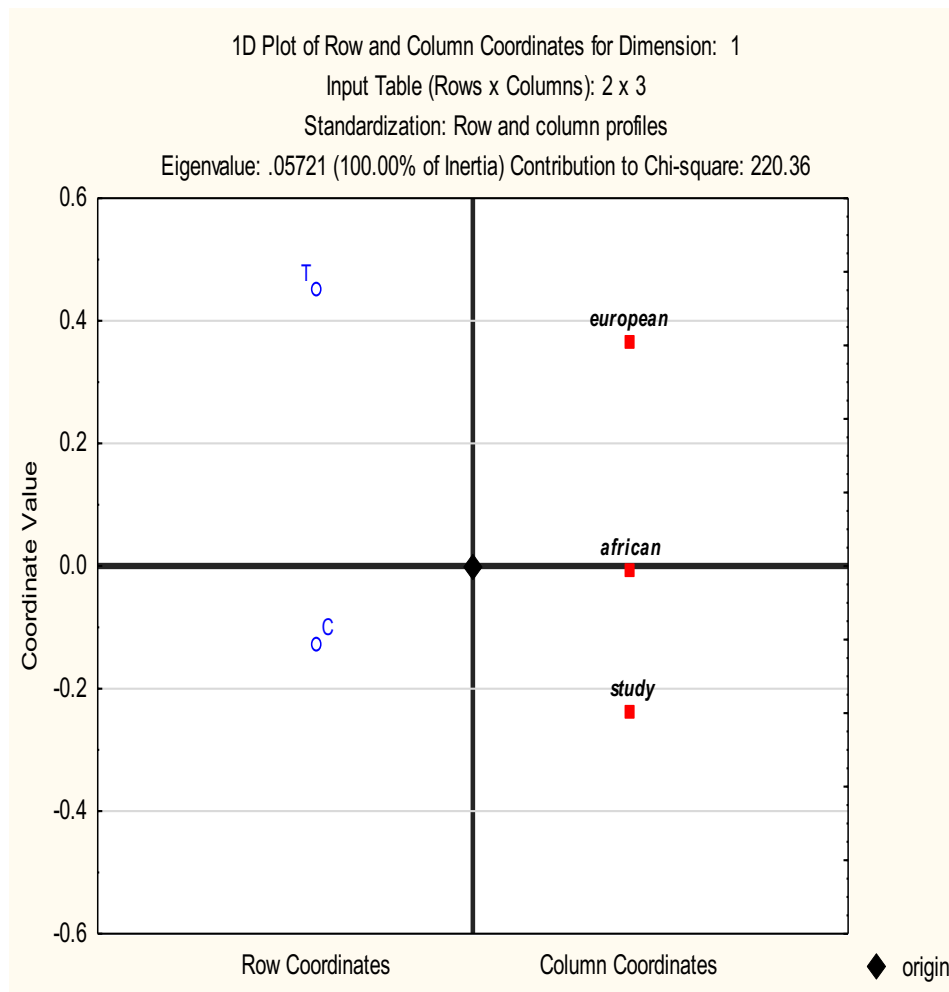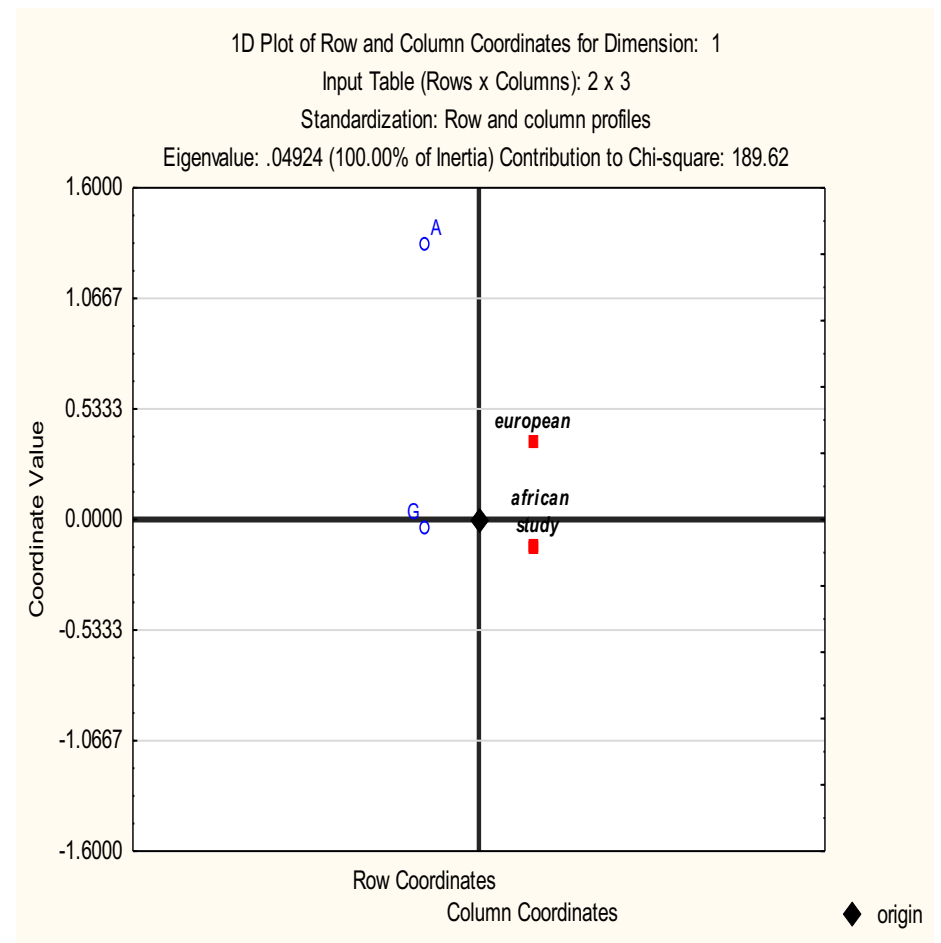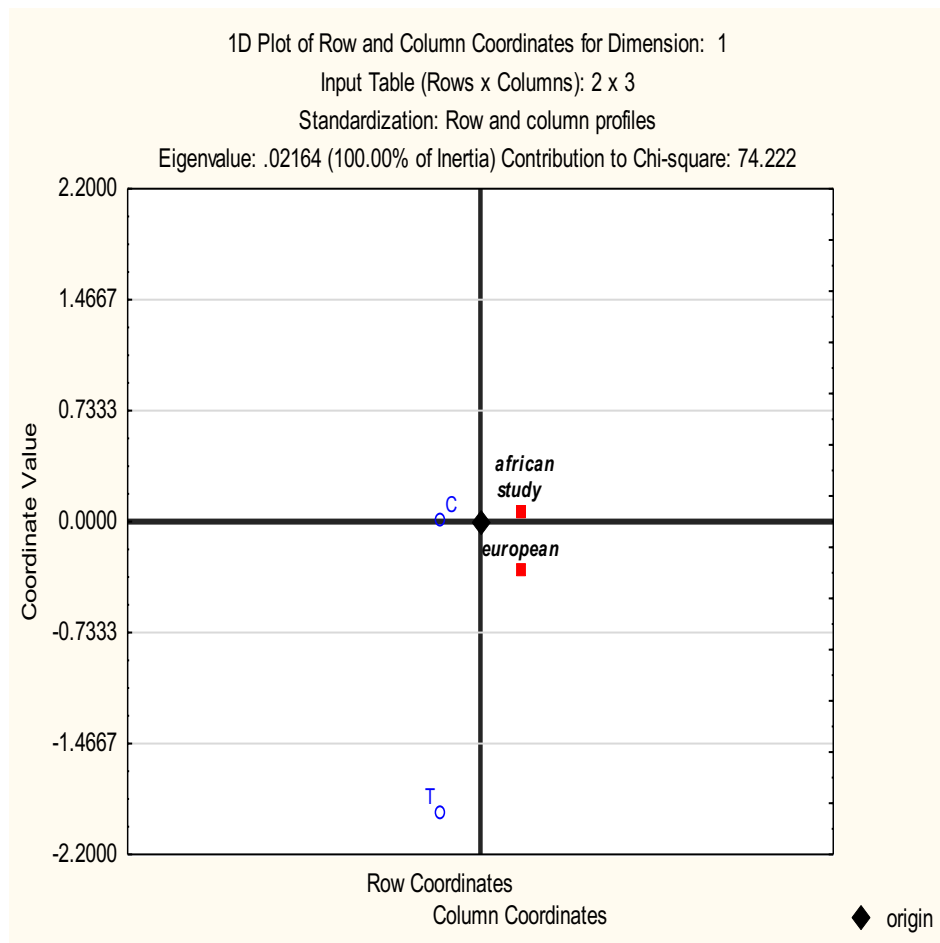**Figure 27.** Correspondence analysis (CA) plot of rs7900194 (*CYP2C9*8*) and rs11045585 (*SLCO1B3*). A closer relationship between the study group and the European group compared to the African group is indicated. The blue circles represent the allele or row coordinates, while the red squares represent the three populations being compared.

Figure 28. Correspondence analysis (CA) plot of rs4646 (*CYP19A1*), rs12248560 (*CYP2C19*17*) and rs11045879 (*SLCO1B1*). The African and European group have a closer relationship to each other than the study group. The blue circles represent the allele or row coordinates, while the red squares represent the three populations being compared.

Figure 29. Correspondence analysis (CA) plot of rs374006 (*ABCC2*) and rs7902557 (*CYP2C19*27*). A distance relationship was detected between all three groups. The blue circles represent the allele or row coordinates, while the red squares represent the three populations being compared.

Figure 30. Correspondence analysis (CA) plot of rs1045642 (*ABCB1*) and rs9024 (*CBR1*). A distance relationship was found between all three groups. The blue circles represent the allele or row coordinates, while the red squares represent the three populations being compared.
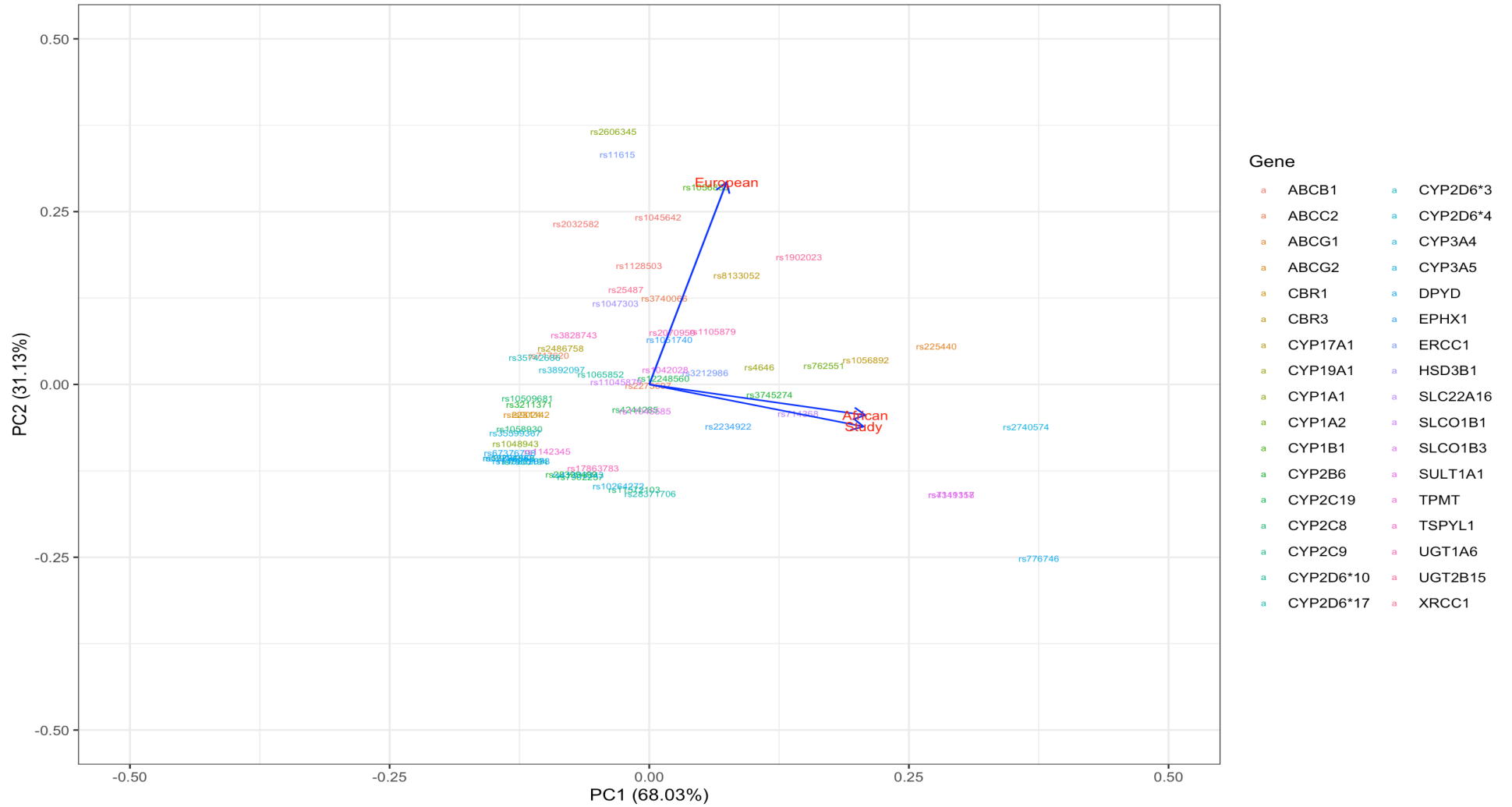
Figure 31. Principle component analysis (PCA), graphical represents the relationship between the study, African and European population. The Study population is more closely related to the African population in comparison to the European population.

4.7 Intra-population variation, through the assessment of linguistic groups

A large portion of the study population (n = 185) classified themselves as Bantu speakers and did not mention their exact mother tongue. These samples were excluded from the intra-population analysis. The three participants that classified themselves as Afrikaans speakers were also excluded.

The TshiVenda speakers are located significantly further North than the Sotho and Tswana speakers. The intra-population assessment was performed, using the Nguni-Tsonga (n = 126), Sotho-Tswana (n = 219) and Venda speakers (n = 231), to assess the intra-population variation (Figure 32).

In Figure 32 the PCA plot depicts the intra-population variability in the SSA population. The Nguni-Tsonga speakers appear to be closer related to the Sotho-Tswana speakers, and the furthest from the Venda speakers. There are some SNPs that appeared to be more prevalent in one particular population than another, for example: rs1056892 (*CBR3)* and rs225440 (*ABCG1*), which was found to be more prevalent in the Sotho-Tswana speakers, and rs776746 (*CYP3A5*3*) and rs2234922 (*EPHX1*), being more prevalent in the Venda speakers. There were also SNPs that showed uniformity between populations, and these included; rs3745274 (*CYP2B6*6*), rs4149117 (*SLCO1B3*), rs7311358 (*SLCO1B3*) and rs4244285 (*CYP2C19*2*), all these SNP cluster close the joining axis. Intra-population variability between African populations of Southern Africa, South/Central Africa, Far West Africa and West Africa, were also reported in the study by da Rocha *et al.*, which focused on ADME genes in the SSA population.[98]

This study further found that the South African region had the highest number novel single nucleotide variations identified.[98] The Venda speakers appeared to be further distanced from other Bantu speakers. The Venda speakers were geographically located further North on the board of Mozambique. This finding further highlights the importance of gaining a better understanding of the variation that exists in ADME genes, particularly in the Sothern African region, as this data is not currently available in genome databases such as the 1000 Genomes Project. The importance of which,

can lead to improved insight into drug efficacy and safety for patients in this region, where disease burden is high and treatment outcome of sub-optimal.
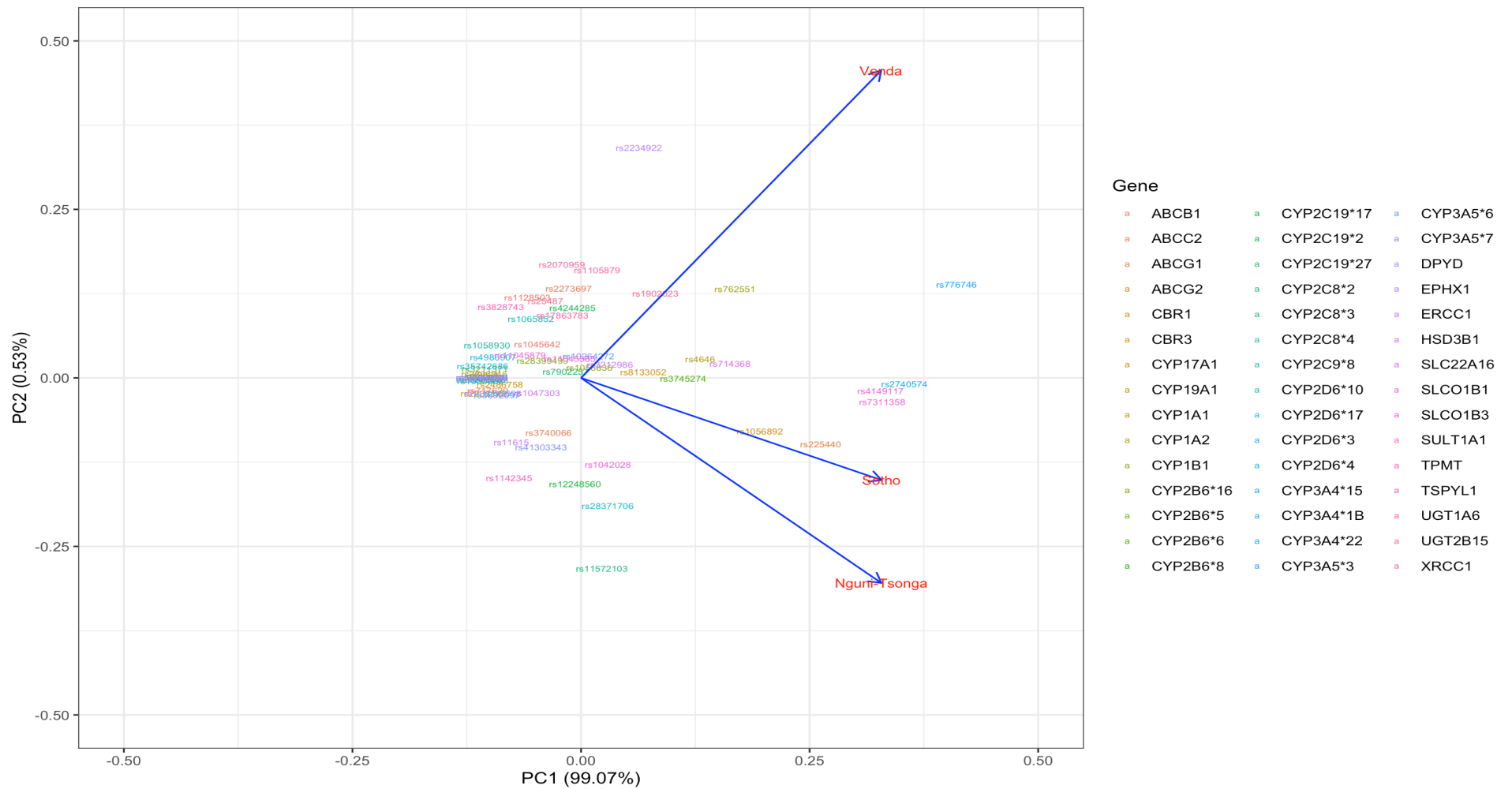
Figure 32. Principle component analysis (PCA), graphical representation of the minor allele frequencies (MAF) between the Nguni-Tsonga, Sotho-Tswana and Venda speakers.

4.8   Identification of the risk allele with associated effects

Genome-wide associated studies (GWASs) have enabled the identification of the relationship between genetic variants and disease.[125] Complex diseases arise from the interplay of multiple genetic and environmental factors.[126] Natural selection has led to a high tendency of risk alleles to be enriched in Mendelian disease. This essentially means that an allele that was previously advantageous or neutral may become harmful, making it a risk allele.[127]

Several types of SNP variants exist and in the current study, the most common were missense variants (60%), followed by intron variants (13.3%) and synonymous variants (8.3%) (Figure 33). This is important to consider when investigating the effect that polymorphisms have on protein function.

Exons are nucleic acid coding sequences, which are present in mRNA, while introns are non-coding sequences and are removed via RNA splicing before translation. Intron variants affect alternative splicing by interfering with the splice recognition site, acting as enhancers for suppressers of genes. However, mutations occurring at the splicing site of introns, can directly affect the exon region directly next to it, resulting in either an insertion, deletion or frameshift mutation. Synonymous or silent variants on the other hand result in nucleotide substitutions that do not change the amino acid. These variants were previously thought not to affect the properties of the synthesise protein, however, the variant can disrupt transcription, splicing, co-translational folding and mRNA stability. Furthermore, they are also largely affected by external factors.[128,129]

In Figure 31 the MAFs of the study population is compared to the MAFs of the African and European populations in the 1000 Genomes Project. In some instances, minor allele was more prevalent in the European population than either of the other two populations, which is also notable on the PCA plot. The minor alleles of several SNPs appeared to be more prevalent in the current study population than that reported for either the African or European population namely; rs225440 (*ABCG1*), rs8133052 (*CBR3*), rs4646 (CYP19A1), rs28399499 (*CYP2B6*16*), rs7902257 (*CYP2C19*17*), rs10264272 (*CYP3A5*6*), rs1051740 (*EPHX1*), rs714368 (*SLC22A16*), rs4149117 (*SLCO1B3*), rs7311358 (*SLCO1B3*) and rs1042028 (*SULT1A1*2).* The implications of

this differences in allele frequency is discussed in section 4.6 below, as it relates to the mechanism of action, and effect on drug metabolism. The population of African ancestry harbours a large pool of genetic variation. Africa is considered the cradle of humankind, with humans inhabiting Africa for > 200 000 years. Variant accumulation and novel variant discovery are the result of evolutionary processes such as random mutation, admixture and genetic drift.[59] Traditionally, African American, Yoruba and Luhya populations have been used to represent populations of African ancestry.[9] However, the data generated in the present study indicates key differences in allele frequencies within the SSA population. The current information on allele frequency is not reflective of all African populations, as most data available has been based on African populations in northern and eastern Africa.
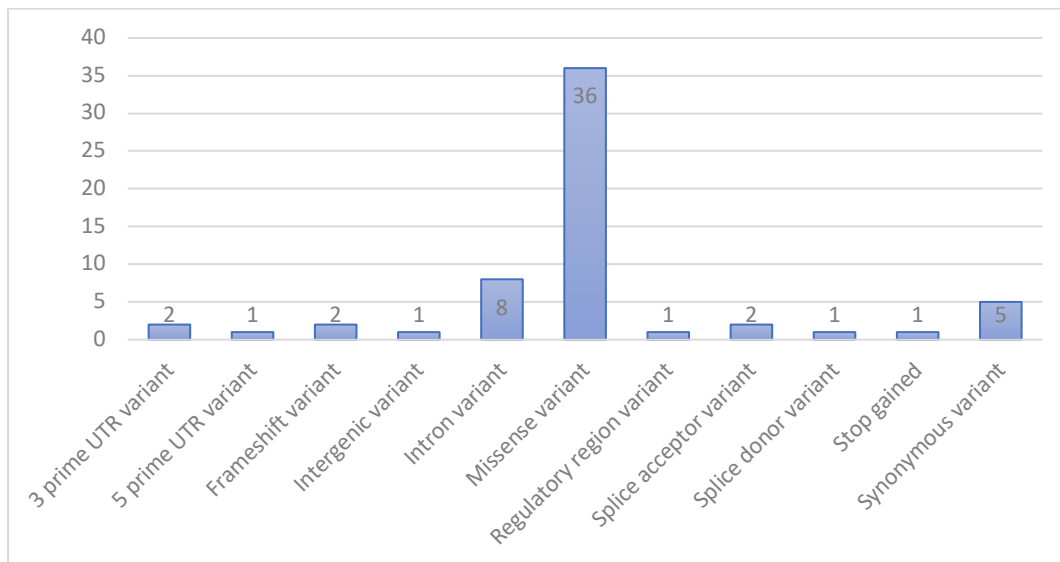
Figure 33. Distribution of variant mutation types investigated in the current study.

Several SNPs were found to be present only in the current study and African population, but absent in the European population; *CYP2B6*16* (rs28399499), *CYP2C19*27* (rs7902257), *CYP2C8*2* (rs11572103), *CYP2D6*17* (rs28371706), *CYP3A4*15* (rs4986907), *CYP3A5*6* (rs10264272) and *CYP3A5*7* (rs41303343). One SNP, *CYP2C9*8* (rs11572103) was absent in the study population but has been reported with a frequency of 0.05 in the 1000 Genomes databases for Africans.

Minor alleles were first described in European populations, as more ethnicities were sequenced and data collected, minor alleles were found to be the ancestral alleles.

The function of the associated protein or enzyme, is the determining factor of risk and whether the allele changes the amino acid encoded subsequently affecting the structure and functional ability of the protein and/or enzyme in-turn affecting the metabolism of the drug (Table 15). For example, the *ABCB1* gene, where both major and minor alleles are regarded as the risk allele, functioning is dependent on the drug, anthracycline or platinum compound, being reviewed. It is important to remember, that the present study only focused on anti-cancer drugs, and the risk alleles may change based on the functional protein produced and the metabolism pathway of a different drug. As our bodies are a network, no single gene works in isolation, therefore one gene could affect multiple drug pathways and the risk allele could vary depending on the drug investigated.

Another example is rs11615 (*ERCC1*), which is essential for the repair of DNA damages cause by cyclophosphamide to both normal and cancerous cells. Asn = rs11615 negatively influences the stability and level of *ERCC1* mRNA and subsequent protein expression. This silent mutation is associated with reduced protein expression and transcript stability. An inability to repair DNA damage is detrimental, however on the other hand the worsted ability to repair DNA in cancer patients has been associated with a better response to genotoxic treatment due to damage and accumulation in cancer cells.[121]

In the present study, focus was placed on the effect of SNPs pertaining to drug metabolism, more specifically anti-cancer drug metabolism, athough it is noted that many of the select SNPs function in the metabolism of other drugs as well. Table 15 summarises the minor alleles and their frequency in the populations, as well as the reported risk allele as it pertains to anti-cancer drug metabolism. The PharmGKB website,[116] was primarily used to establish the risk alleles, as well as the level of significance and overall functional effect of the polymorphisms. The PharmGKB database is the result of articles published worldwide. Information pertaining to drug dosage or dosage adjustment recommendation, can be sourced from published data on the Clinical Pharmacogenetics Implementation Consortium (CPIC).[130]

For example; rapid metabolisers may benefit from a higher dosage, while poor metabolisers may benefit from a lower dose to reduce the potential for ADRs.[131]

Clinical annotations in terms of the relationship between a gene and drug is graded according to a level of evidence, and ranked between 1 and 4 (1A, 1B, 2A, 2B, 3 and 4). PharmGKB level 1A is the highest annotation level, where the variant-drug combination is included in the CPIC.[116] Level 1B, is indicative of an association that has been replicated in more than one cohort, with a significant *P*-value. Both of which have a large body of evidence showing the association of altered drug response. Level 2, includes variants of moderate evidence, while in levels 3 and 4 the evidence is not yet strong enough to be used for clinical translation.[116] It is important to note, that the absence of level 1 annotation, in some instance is likely as a result of a lack of evidence and associated studies, as it relates to African-specific variants, rather than evidence against clinical relevance.

Table 15. Summary of minor allele frequencies (MAF) and risk alleles.

dbSNP	Gene	Variant type	MAF Allele	Ave. MAF	Study	AFR	EUR	PharmG Kb[132]	Level	Drug	Effect
rs2032582	ABCB1	Missense	T	0,33	0,00	0,02	0,47	T / C	3	Platinum compounds / taxanes	Efficacy / Toxicity
rs1045642	ABCB1	Missense	A	0,40	0,11	0,15	0,52	G / A	3	Anthracyclines, methotrexate, platinum compounds, paclitaxel	Efficacy / Toxicity
rs1128503	ABCB1	Synonymous	A	0,42	0,10	0,14	0,42	A / G	3	Paclitaxel and platinum compounds	Toxicity/ADR
rs3740066	ABCC2	Missense	T	0,29	0,12	0,22	0,37	T	3	FAC, methotrexate, platinum compounds	Toxicity/ADR
rs2273697	ABCC2	Missense	A	0,19	0,17	0,19	0,20	G	3	FAC, methotrexate, platinum compounds	Toxicity/ADR
rs717620	ABCC2	5'-UTR	T	0,14	0,01	0,03	0,21	T	3	5-FU, leucovorin, oxaliplatin	Toxicity/ADR
rs225440	ABCG1	Intronic	T	0,43	0,60	0,59	0,41	T	3	5-FU, leucovorin, oxaliplatin	Toxicity/ADR
rs2231142	ABCG2	Missense	T	0,12	0,01	0,01	0,09	T	3	FAC, methotrexate	Toxicity/ADR
rs9024	CBR1	3'-UTR	A	0,13	0,01	0,01	0,09	G	3	Anthracyclines	Toxicity/ADR
rs8133052	CBR3	Missense	A	0,37	0,29	0,25	0,45	G	3	Anthracyclines	Efficacy / Toxicity
rs1056892	CBR3	Missense	A	0,43	0,48	0,51	0,35	G	3	Anthracyclines	Toxicity/ADR
rs2486758	CYP17A1	Regulatory region	C	0,18	0,03	0,04	0,23	C	N/A	Androgen deprivation therapy response	Efficacy
rs4646	CYP19A1	3'-UTR	A	0,34	0,40	0,28	0,29	C	3	FAC, docetaxel, epirubicin, paciltaxel, TAM	Efficacy
rs1048943	CYP1A1	Missense	C	0,13	0,00	0,01	0,03	C	3	Capecitabine/ docetaxel	Efficacy
rs2606345	CYP1A1	Intronic	A	0,28	0,01	0,05	0,66	A	3	Carbamazepine, phenobarbital, phenytoin, valproic acid	Efficacy
rs762551	CYP1A2	Intronic	C	0,37	0,43	0,44	0,32	C	3	Imatini	Toxicity & dose adjustment
rs1056836	CYP1B1	Missense	G	0,48	0,20	0,18	0,6	C	3	FAC, epirubicin	Efficacy & Toxicity/ADR
rs12721655	CYP2B6*8	Missense	G	<0,01	0,00	0,00	0,00	G	3	Cyclophosphamide / doxorubicin	Efficacy
rs3211371	CYP2B6*5	Missense	T	0,05	0,01	0,01	0,11	T	3	Cyclophosphamide / doxorubicin	Toxicity/ADR
rs28399499	CYP2B6*16	Missense	C	0,02	0,12	0,08	0,00	C	3	Nevirapine	Toxicity
rs3745274	CYP2B6*6	Missense	T	0,32	0,36	0,37	0,24	T	3	Efavirenz	Dosage & Toxicity
rs4244285	CYP2C19*2	Synonymous	A	0,22	0,17	0,17	0,15	A	3	Cyclophosphamide / doxorubicin	Efficacy
rs12248560	CYP2C19*17	Intronic	T	0,15	0,16	0,24	0,22	C	3	FAC, tamoxifen	Efficacy & Toxicity/ADR
rs7902257	CYP2C19*27	Intronic	A	0,15	0,15	0,08	0,00	A	N/A	Multiple drugs	Decreased function /Efficacy/Toxicity

dbSNP	Gene	Variant type	MAF Allele	Ave. MAF	Study	AFR	EUR	PharmGKb[132]	Level	Drug	Effect
rs10509681	CYP2C8*3	Missense	C	0,05	0,00	0,01	0,12	C	3	Paclitaxel	Metabolism/PK
rs1058930	CYP2C8*4	Missense	C	0,02	0,01	0,00	0,06	C / G	3	Paclitaxel, platinum compound	Toxicity
rs11572103	CYP2C8*2	Missense	A	0,05	0,21	0,19	0,00	A	3	Paclitaxel	Toxicity
rs7900194	CYP2C9*8	Missense	T	0,01	0,00	0,05	0,00	A	3	Warfarin	Response
rs1065852	CYP2D6*10	Missense	A	0,24	0,10	0,11	0,2	A	1A	Tamoxifen	Metabolism/PK
rs3892097	CYP2D6*4	Splice acceptor	T	0,09	0,03	0,06	0,19	T	1A	Tamoxifen	Efficacy & Toxicity
rs35742686	CYP2D6*3	Frameshift	-	0,01	0,00	0,00	0,2	Deletion	1A	Tamoxifen	Efficacy/Metabolism/PK
rs28371706	CYP2D6*17	Missense	A	0,06	0,23	0,22	0,00	A	1A	Tamoxifen	Efficacy/Metabolism/PK
rs4986907	CYP3A4*15	Missense variant	T	0,01	0,02	0,02	0,00	T	N/A	Multiple drugs	Loss of function – Efficacy/Metabolism
rs2740574	CYP3A4*1B	Intergenic	C	0,23	0,75	0,77	0,3	C	3	Docetaxel, tamoxifen	Toxicity/ADR & Metabolism/PK
rs35599367	CYP3A4*22	Intronic	A	0,01	0,002	0,00	0,05	C	3	Odansetron	Metabolism/PK
rs776746	CYP3A5*3	Splice acceptor	T	0,37	0,84	0,82	0,06	T	3	Paclitaxel	Toxicity/ADR
rs10264272	CYP3A5*6	Synonymous	T	0,04	0,20	0,15	0,00	T	3	Odansetron, cabazitaxel, paclitaxel, carboplatin, docetaxel	Toxicity/ADR
rs41303343	CYP3A5*7	Frameshift	A	0,03	0,10	0,12	0,00	A	3	Odansetron, cabazitaxel, paclitaxel, carboplatin, docetaxel	Toxicity/ADR
rs55886062	DPYD	Missense	C	<0,01	0,00	0,00	0,00	C	3	FAC	Toxicity/ADR & Metabolism
rs67376798	DPYD	Missense v	A	<0,01	0,00	0,00	0,01	A	3	FAC	Toxicity/ADR & Metabolism
rs3918290	DPYD	Splice donor	T	<0,01	0,00	0,00	0,00	T	3	FAC	Toxicity/ADR & Metabolism
rs115232898	DPYD	Missense	C	0,01	0,02	0,02	0,00	C	1A	5-FU	Toxicity/ADR
rs2234922	EPHX1	Missense	G	0,22	0,28	0,35	0,16	G	3	Docetaxel	Dosage
rs1051740	EPHX1	Missense	C	0,31	0,25	0,14	0,30	C	3	Cisplatin, cyclophosphamide	Toxicity
rs3212986	ERCC1	Stop gained	A	0,30	0,23	0,29	0,25	A	3	Platinum compound response	Toxicity
rs11615	ERCC1	Synonymous variant	A	0,30	0,05	0,04	0,62	A / G	3	FAC, platinum compound response	Efficacy/Toxicity
rs1047303	HSD3B1	Missense	C	0,17	0,11	0,09	0,34	C		Androgen deprivation therapy response	Efficacy
rs714368	SLC22A16	Missense	C	0,31	0,44	0,39	0,22	T	3	FAC	Efficacy/Toxicity
rs11045879	SLCO1B1	Intron variant	C	0,22	0,07	0,19	0,19	C	3	Methatrexate response	Toxicity

dbSNP	Gene	Variant type	MAF Allele	Ave. MAF	Study	AFR	EUR	PharmG Kb[132]	Level	Drug	Effect
rs4149117	SLCO1B3	Missense	T	0,30	0,71	0,64	0,14	G	3	Carboplatin, paclitaxel	Toxicity
rs7311358	SLCO1B3	Missense	G	0,30	0,71	0,64	0,14	A	3	Carboplatin, paclitaxel	Toxicity
rs11045585	SLCO1B3	Intronic	G	0,16	0,15	0,22	0,15	G	3	Docetaxel	Toxicity
rs1042028	SULT1A1	Missense	T	0,43	0,23	0,17	0,24	T	3	Tamoxifen	Efficacy
rs1142345	TPMT	Missense	C	0,04	0,04	0,07	0,03	C	3	Cisplatin, cyclophosphamide	Toxicity/ADR
rs3828743	TSPYL1	Missense	A	0,35	0,04	0,06	0,26	A	3	Abiraterone / prednisolone	Efficacy
rs2070959	UGT1A6	Missense	G	0,28	0,15	0,24	0,31	A	3	Irinotecan	Toxicity/ADR
rs17863783	UGT1A6	Synonymous	T	0,07	0,13	0,13	0,02	T	4	Anthracyclines	Toxicity/ADR
rs1105879	UGT1A6	Missense	C	0,32	0,21	0,30	0,33	A	3	Irinotecan	Toxicity/ADR
rs1902023	UGT2B15	Missense	A	0,45	0,31	0,40	0,51	C	3	Tamoxifen	Toxicity/ADR & Survival
rs25487	XRCC1	Missense	T	0,26	0,11	0,11	0,37	T	2B	Platinum compound response	Efficacy / Toxicity

* Red letters, reflect the risk alleles that differ from the minor allele according to data obtained from PharmGKB.
* ADR: Adverse drug reaction, PK: Pharmacokinetics

4.9 Clinical annotation and pharmacogenetic implication

This section focuses on various individual anti-cancer drugs and the genes involved in the metabolism of the drugs. Each polymorphism affecting the metabolism of the drug will be discussed both individually, as well as in conjunction with each other. The PharmGKB[132] and CPIC[130] databases were used to compile the information provided.

4.9.1 Androgen deprivation therapy / Abiraterone

Abiraterone acetate is a prodrug that is converted to the active metabolite abiraterone via esterase-catalysed hydrolysis.[120] Abiraterone is a selective inhibitor of 17α-hydroxylase and C17,20-lyase enzymatic activity of CYP17. Androgen receptor signalling is essential in the progression from primary to metastatic PrCa, and CYP17 is required for androgen biosynthesis. Table 16 lists all of the genetic variants that are associated with the metabolism of abiraterone, that were investigated in the study. The C allele has been associated with a shorter time to biochemical progression and biochemical response, due to the upregulation of CYP17A1 activity, resulting in a poorer response to abiraterone.[133] The C allele of *CYP17A1* was shown to have a lower prevalence in the current study (0.03), as well as the African population (0.04) in comparison to the European population (0.23). Another genetic polymorphism that has been linked to abiraterone response is *HSD3B1* (rs1047303), where the C allele has been associated with a decreased response to abiraterone. The allele change 1245A→C, results in an amino acid change of the Asn→Thr, rendering the 3β-hydroxysteroid dehydrogenase-1 enzyme resistant to proteasomal degradation, which leads to an accumulation of this enzyme and a gain of function.[72] The C allele was found to have an allele frequency of 0.11 in the study population, and 0.09 in the African population, whereas it is dominant in the European population (0.34).

TSPYLs are involved in many cellular functions and regulation of gene expression. Genetic polymorphisms, including mutations and methylation status changes affect the regulatory effect of the *TSPYL* genes.[73] The *CYP3A4* gene is abundant in the liver and is responsible for the conversion of abiraterone to the inactive metabolite N-oxide abiraterone sulfate, which is then excreted via the urine. The *TSPYL* gene induces *CYP17A1* expression but suppresses *CYP3A4 expression*, enabling the active

106

metabolite to circulate for longer in the system and effectively inhibiting *CYP17A1* expression.[73] However, the rs32828743 (*TSPYL1)* polymorphism, particularly the A allele, abolishes TSPYL's ability thereby suppressing the expression of *CYP3A4*, resulting in an increased expression and decreased abiraterone exposure, consequently result in an increased synthesis of androgen. The A allele was found to have a low prevalence in the study population (0.04). The *CYP3A4* gene is also involved in the conversion of abiraterone to its inactive metabolite, polymorphisms within the *CYP3A4* gene which alter the enzyme activity, cause a decrease in the conversion of active metabolites to inactive metabolites, which increases the exposure of *CYP17A1* to abiraterone.[73] The associated risk allele for *CYP3A4*1B* polymorphism indicated the highest prevalence of 0.75, the C- allele is also the ancestral allele and common in the African population. While the risk allele for CYP3A4*15 and *22, show a prevalence of 0.02 and 0.002 respectively. In this event, the CYP3A4 polymorphism has a beneficial role to play, as the exposure of the active metabolite is prolonged. *SULT1A1* polymorphism, was one of the six SNPs that deviated from HWE, therefore no conclusion could be drawn from this SNP.

Table 16. Summary of polymorphisms affecting the metabolism of androgen deprivation therapy / abiraterone and allele frequency of the study cohort of the identified risk allele.

dbSNP	Gene (star allele)	Risk allele	Study allele Freq.	Details
rs1047303	*HSD3B1*	C	0.11	C - associated with 3β-hydroxysteroid dehydrogenase-1 resistance to proteasomal degradation, causing substantial accumulation of this enzyme and gain of function.[72]
rs2486758	*CYP17A1*	C	0.03	C - associated with the upregulation of CYP17A1 activity, resulting in a poorer response to AA/P treatment.[134]
rs3828743	*TSPYL1*	A	0.04	A - associated with decreased response to AA/P treatment, involved in the regulation of CYP3A4.[73]
rs4986907	*CYP3A4(*15)*	T	0.02	These alleles are associated with a decreased activity. CYP3A4 in conjunction with SULT1A1 are involved in the metabolism of abiraterone to inactive metabolites. [73]
rs2740574	*CYP3A4(*1B)*	C	0.75	
rs35599367	*CYP3A4(*22)*	T	0.002	

NCBI: National Center for Biotechnology Information
*AA/P: Abiraterone acetate and prednisolone

Although there are clinical guidelines available on the CPIC for abiraterone, there is limited literature available on genetic variants affecting the efficacy of abiraterone in men with PrCa. There is also little to no literature available regarding the penetrance of these genes.

4.9.2 Platinum compounds

Platinum-based drugs are the largest class of drugs used to treat cancer, and used in the treatment of both BrCa and PrCa. They destroy cancer cells by interfering with the DNA. This interference occurs through inter- and intra-strand crosslinks and DNA-protein crosslinks, thereby preventing cell division and growth.[135] The influx and efflux of platinum drugs are regulated by several transport proteins, including ABCC2 and ABCG2. Polymorphisms occurring within transport genes can affect the efficacy and toxicity of a platinum compound. Polymorphisms affecting multidrug resistance transport genes in particular, change the function of the P-glycoprotein (P-gp) which extrude through the cell membrane, and regulates the transport of molecules in and out of the cell.[135] Depending on the function of the transport genes, they may affect the level of platinum entering the cell, which affects the response to platinum based treatment (efficacy). Alternatively, they can affect the export of the compound out of the cell, leading to accumulation of the compound which could cause toxicity and ADRs. The risk allele T for rs2032582 was absent in the study population, therefore posing little to no risk. The prevalence of the risk allele A for rs1045642 was found to be 0.11, which is significantly lower than in the European population (0.52). Given the elevated risk of the genetic polymorphism, rs1228503 genetic screening prior to treatment initiation with combination therapy of cisplatin, cyclophosphamide, doxorubicin and methotrexate should be performed.

Nucleotide excision repair is controlled by XRCC1, ERCC1 and ERCC2. Known genetics variants in these genes have previously been associated with platinum-based drug response. *XRCC1* (rs25487) is a well-studied nucleotide and has a 2B level of evidence in terms of treatment prediction. The T allele, which results in a substitution of an arginine for glutamine, affects secondary structure features, which are critical for the accurate protein-protein interactions in the BRCT1 domain. This compromises the

DNA repair capabilities, and decreases the response to treatment.[136,137] The prevalence of the T-allele is 0.11 in the study population, a higher prevalence has been reported in the European population (0.37). The importance of screening for this SNP is highlighted by the evidence level of 2B and the evidence score of 11.25, as per PharmGKB. There has been a total of 19 publications, 14 of which have provided a positive association between the T allele of rs25487 and treatment response. Table 17 provides a summary of all the polymorphisms associated with the metabolism of platinum compounds. According to CPIC's gene-drug pairs ERCC1 gene classified as a level D interaction. [130]

Table 17. Summary of polymorphisms affecting the metabolism of platinum compounds and allele frequency of the study cohort of the identified risk allele.

dbSNP	Gene (star allele)	Risk allele	Study allele Freq.	Details
rs2032582	*ABCB1*	T	0	T - associated with an increased risk for gastrointestinal toxicity.[135]
rs1045642	*ABCB1*	A	0.11	A - associated with a poorer response to platinum-based chemotherapy. [135]
rs1128503	*ABCB1*	G	0.90	G - associated with an increased risk of death when treated with a combination of cisplatin, cyclophosphamide, doxorubicin and methotrexate.[135]
rs717620	*ABCC2*	T	0.01	T - associated with neurotoxicity syndrome when treated with FOLFOX, it has also been linked to decreased risk of thrombocytopenia.[138]
rs1058930	*CYP2C8(*4)*	G	0.99	G - associated with increased severity of thrombocytopenia when treated with carboplatin and gemcitabine.[139]
rs11572103	*CYP2C8(*2)*	Del/del	0.79	Persons with the del/del genotype may experience an increased severity of thrombocytopenia when treated with carboplatin and gemcitabine.[139]
rs1051740	*EPHX1*	C	0.25	C - associated with grade 1-4 nephrotoxicity.[140]
rs3212986	*ERCC1*	C	0.77	C - increased risk for nephrotoxicity.[140]
rs4149117	*SLCO1B3*	C	0.29	G - linked to an increased risk for anaemia.[141]
rs7311358	*SLCO1B3*	A	0.71	A - linked to an increased risk for anaemia.[141]
rs25487	*XRCC1*	T	0.11	T - associated with a decreased response to treatment when treated with platinum compounds.[136,137]

NCBI: National Center for Biotechnology Information
FOLFOX: Fluorouracil, leucovorin, oxiplatin

Unfortunately, no conclusion could be drawn from *ERCC1* (rs11615), as the polymorphism did not satisfy the HWE test. The G allele of this SNP has been associated with decreased survival in patients treated with cisplatin-based chemotherapy.[140] It is important to note that other genetic and clinical factors may also

influence an individual's response to a specific therapy regime. Given the high prevalence of the genetic variants, population screening prior to treatment may be beneficial. Currently there is no functional evidence of this variation, thus the penetrance of the variant is not available.

Another factor which is understudied in disease treatment is the clinicopathological characteristics of pathogenic variants. In diseases such as BrCa, the gene causing pathogenic variation in conjunction with clinical presentation of the cancer, play a role in treatment determination. *BRCA* is the most widely studied pathogenetic variant, with more than 1600 variants known for *BRCA1* and 1800 for *BRCA2*. Histological types for *BRCA1* pathogenetic variant BrCa include invasive ductal carcinomas and atypical medullary. Triple negative BrCa occurs in 66-100% of *BRCA1* pathogenetic variant BrCas, and in 14-35% of *BRCA2*. *BRCA* pathogenic variants should be taken into account when selecting chemotherapy treatment. *BRCA1* pathogenic variants are more sensitive to platinum-based chemotherapy and more resistant to microtubule-inhibiting chemotherapy such as taxanes. Poly (adenocine diphosphate [ADP]-ribose) polymerase (PARP) inhibitors are a new set of chemotherapy drugs in various phases of FDA approval. The various PARP inhibitors are undergoing drug trials focused on the clinical characteristics and histological grading of the cancer.[142] This is important when deciding to conduct DNA testing for translational purposes.

A genetic variation of interest linked to an ototoxic effect in adult patients with PrCa is *SLC16A5* (rs4788863) when treated with cisplatin. As cisplatin induces the expression of SLC16A5, patients were less likely to experience an ototoxic effect. Unfortunately, the present study did not assess the prevalence of this SNP, however, from the 1000 Genomes Project the C allele indicated a prevalence of 0.729 in the African population, and is protective against ototoxicity.[143]

4.9.3 FAC protocol

The FAC protocol is a chemotherapy regimen that includes the drugs; fluorouracil, doxorubicin and cyclophosphamide. It is the most common BrCa chemotherapy. Given

the complex nature of the pathways of each FAC drug, it is expected that the mechanisms causing the development of ADRs to these three drugs are complex as well.[144] Several functional variants in gene encoding proteins are involved in FAC transport, metabolism and drug-induced damage repair, and these are depicted in Table 18.[144] The *DPYD* gene is of clinical importance for the metabolism of 5-FU. Three of the polymorphisms investigated in the present study were found to be fixed within the study population, and have been discussed previously. Nucleotide variation, rs1152323898 (*DPYD*), which is only present in the African population (0.02), and absent in the European population, is an essential SNP for genetic screening prior to treatment initiation. The mutation carries the phenotype of a poor metaboliser with an activity score of 0.5. All the other DPYD (rs55886062, rs67376798 and rs3918290) were homozygote for the wild-type allele, indicative of a normal metabolizer. According to the CPIC guideline, CPIC level A and PharmGKB level rating of 1A, the guidelines are actionable. The current guidelines suggest a reduction in the dose of 5-FU by 50% for intermediate metabolisers with an activity score of 1.5. The latter refers to heterozygous carriers of decreased function variants, rs67376798, rs75017182 and rs56038477, or homozygous carriers for rs67376798 with an activity score of 1.[130]

Transport proteins, as mentioned previously, regulate the transport of the drug into and out of the cell, either the cancer cell or the liver cell. ATP-cassette transporters are of particular importance and include; ABCB1, ABCC1 ABCC2, ABCG2.[145] Variations in these genes have been linked to various toxicities and ADRs, particularly anaemia.[145] The *SLC22A16* gene is responsible for the transport of the various FAC compounds into the cell. Polymorphisms associated with this transport gene affect the transport of FAC molecules into the cell. Doxorubicin is metabolized by the CBR enzymes into its active component. Cyclophosphamide, also requires activation by a number of different CYP2C family genes.

ERCC1 is essential for the repair of DNA damage in the cells caused by cyclophosphamides. The polymorphisms, rs11615 and rs3212986, are silent variants and have been linked to lower protein expression, reduced transcript stability and protein levels.[145] Lowered expression of this gene would result in DNA repair not taking place, and accumulation of damage in cancer cells. Normal cells also suffer from this treatment due to low levels of DNA repair.[145] XRCC1 on the other hand plays a key

role in the BER pathway, which is involved in DNA repair, after exposure to single strand breaks. The prevalence of the rs25487-C allele is significantly higher in the African population (0.89). Evidence of leukopenia and recurrent neutropenia, has been linked to the genes' ability to metabolize the substrates of the FAC protocol.

*CYP2C19*17* increases the transcriptional activity, essentially resulting in an ultra-metaboliser. The C allele was highly prevalent in the study population (0.84) and should form part of the screening process.[145] Conversely, the *CYP2C19*2* variant has aberrant mRNA splicing, resulting in the absence of hepatic enzyme and a poor metaboliser phenotype.[145] Poor metabolisers are unable to break the drug down into various metabolites, resulting in drug accumulation which may cause drug toxicity. A reduction in elimination rate also results in the accumulation of compounds, which could also lead to toxicity. Another SNP, which should be added to the screening panel, is that of CYP2B6*6, as it plays a role in multiple drugs. It was found to have a high prevalence in the study population (0.64). Anaemia is the most common side effect and has been linked to all three functional gene groups, DNA repair, metabolism and transport. Other effects include gastrointestinal toxicity, nausea and vomiting.[145] Adverse effects during FAC therapy are a complex interplay among polygenic inheritance and clinical risk factors. In order to reduce the side effects and potential FAC toxicities, genetic screening prior to treatment is advised. Patients at high risk for chemotherapy-induced anaemia, could be supplemented with iron, folic acid and B6 vitamin during the initial stage of treatment, rather than when presenting with ADRs. Dosage monitoring and adjustments is the future for personalize approach to cancer treatment, tailoring treatment to improve patient outcomes and quality of life.

The anti-cancer drug class of taxanes includes; cabazitaxel, docetaxel and paclitaxel. These agents are used for the treatment of various cancers including advanced or metastatic BrCa and metastatic PrCa. These drugs interfere with microtubule growth. Taxanes block cell division by binding to the alpha tubulin which leads to mitotic arrest and cell death.[146] Common toxicities associated with this drug class includes; neuropathy, anaemia and gastrointestinal complications.[146] Metabolising enzymes involved in the metabolism of paclitaxel include, CYP3A4 and CYP2C8. CYP3A5 and CYP3A4 are involved in the metabolism of docetaxel. Both drugs are substrates for

the ATP binding cassette of multidrug transporters; ABCB1, ABCG2, ABCC1 and ABCC2. Additionally, SLCO1B3 is the most efficient influx transporter for docetaxel.

Table 18. Summary of polymorphisms affecting the metabolism of FAC protocol drugs and allele frequency of the study cohort of the identified risk allele.

dbSNP	Gene (star allele)	Risk Allele	Study allele Freq.	Details
rs1045642	ABCB1	G	0.89	G - associated with an increased risk of anaemia when treated with the FAC protocol.[145]
rs3740066	ABCC2	T	0.12	T - associated with an increased risk of nausea and neutropenia when treated with FAC.[145]
rs2273697	ABCC2	G	0.83	G - associated with an increased risk of anaemia when treated with cyclophosphamide, doxorubicin and fluorouracil (FAC).[145]
rs2231142	ABCG2	T	0.01	T - associated with an increased risk of anaemia.[145]
rs12721655	CYP2B6(*8)	G	0	G - linked to a decreased survival when treated with cyclophosphamide and doxorubicin in women with breast neoplasms.[145]
rs3211371	CYP2B6(*5)	T	0.01	T - increased likelihood of dose delay when treated with cyclophosphamide and doxorubicin in women with breast neoplasms.[145]
rs3745274	CYP2B6(*6)	G	0.64	The G allele carriers are more likely to require a reduction in dose.[145]
rs12248560	CYP2C19(*17)	C	0.84	C - associated with an increased risk for leukopenia when treated with the FAC protocol.[145]
rs4244285	CYP2C19(*2)	A	0.17	A - associated with an increased risk of neutropenia when treatment with the FAC.[145]
rs3212986	ERCC1	A	0.23	A - associated with an increased risk of neutropenia when treated with FAC[145]
rs1056836	CYP1B1	C	0.80	C - associated with an increased risk for nausea when treated with FAC.[145]
rs714368	SLC22A16	T	0.56	T - may have decreased risk for nausea, but an increased likelihood of dose delay.[145]
rs25487	XRCC1	C	0.89	C - associated with increased risk for nausea, but a decreased likelihood of dose delay[145]

NCBI: National Center for Biotechnology Information

A cassette of multidrug transporters are responsible for the efflux of taxanes from the cell. Mutations in these transport genes can affect the response to treatment and increase the risk of toxicity. These polymorphisms affect the P-gp function, which may increase toxicity and reduce efficacy.[147] ABCB1 variant rs2032582 – C causes a wobble effect on the amino acid glycine that is produced, resulting in decreased clearance of paclitaxel. This variant appeared to be fixed in the study population with

a prevalence of 0.98 in the African population and 0.53 in the European population. Variant rs1045642 – A, an amino acid substitution of an alanine to serine or threonine, had a much lower prevalence in the study population (0.11), but higher prevalence in the European population (0.52). Changes in this nucleotide increases the efflux of antineoplastic agents from the cancer cells, resulting in lower plasma concentrations of the drug and thus reduced therapeutic efficacy.[148] Previous studies conducted within the African population focusing on genetic diversity of transporter genes, found varied results of the effect of the ABCB1 transport variants rs2033582, rs1045642 and rs1128503. Some studies noted no effect of the subtract drugs Nevirapine and Rifampin, HIV and TB drugs respectively, where other studies showed significant difference in Efavirenz concentration between genotypes.[96] This echos the effect the gene has on the phenotype, and the mechanism of action of the drug and the role the gene plays in the transport of the drug within the body.

4.9.4 Taxanes

SLCO1B3 is a solute organic protein carrier and is responsible for the uptake of both docetaxel and paclitaxel.[141] Mutations in this gene result in lower levels of the drug in circulation and a higher AUC, which has been associated with toxicity. There is a higher risk for anaemia when treated with a combination of carboplatin and paclitaxel.[141] The risk allele of both polymorphisms were found to occur with a frequency of 0.29 in the study population, which is lower than the European population (0.86).

Paclitaxel is hydroxylated in the liver by CYP2C8. Variants that affect the activity of the CYP2C8 enzyme, result in decreased hydroxylation of paclitaxel. This potentially causes toxicity or treatment failure.[149] The *1/*2 allele of CYP3A4*1, *8 and *20 (although not part of the current panel of selected SNPs), has been linked to an increased risk for neuropathy when treated with paclitaxel.[149] CYP3A4 and CYP3A5 are involved in the metabolism of multiple other drugs.[149] CYP3A5*3 rs776746 – T allele is associated with a complete loss of function cause by the alternative splicing by the variant. The prevalence was significantly higher in the study cohort and African population, 0.84 and 0.82, respectively, compared to the European population (0.06). Similar ratios were seen in the high coverage African dataset, in a study conducted by

de Rocha *et al.*[144] Routine monitoring of this polymorphism is thus critical. Table 19 summarises the genes involved in the metabolism of taxanes, and associated allele frequencies in the study cohort.

Table 19. Summary of polymorphisms affecting the metabolism of taxanes and allele frequency of the study cohort of the identified risk allele.

dbSNP	Gene (star allele)	Risk Allele	Study allele Freq.	Details
rs2032582	*ABCB1*	C	1	C - associated with a decreased response to paclitaxel.[147]
rs1045642	*ABCB1*	A	0.11	A - associated with an increased risk of neuropathy when treated with paclitaxel.[150]
rs10509681	*CYP2C8(*3)*	C	0	Both polymorphisms have been associated with decreased enzyme activity, resulting in the decreased metabolism of paclitaxel.[149]
rs1058930	*CYP2C8(*4)*	C	0.01	
rs776746	*CYP3A5(*3)*	T	0.84	T - associated with an increased risk of neurotoxicity when treated with paclitaxel.[151]
rs4149117	*SLCO1B3*	G	0.29	G - associated with an increased risk of anaemia and thrombocytopenia when treated with carboplatin and paclitaxel.[152]
rs7311358	*SLCO1B3*	A	0.29	A - associated with an increased risk of anaemia when treated with carboplatin and paclitaxel.[141]

NCBI: National Center for Biotechnology Information

4.9.5 Anthracyclines

Anthracyclines have an intercalating function and are inserted between adjacent base pairs of DNA, inhibiting DNA and RNA synthesis, particularly in high replicating cells, thereby blocking cell division.[85] Doxorubicin, is an example of an anthracycline. The *SLC22A16* gene is responsible for the transport of anthracycline into the cancer cell. Genetic variation in transporter genes have been associated with drug resistance. The T allele of the polymorphism rs714368, has been associated with a decreased exposure to doxorubicin, as a result polymorphisms affect the transport of the drug into the cancer cell.[153] The T allele was found to have a prevalence of 0.56 in the study population and 0.78 in the European population. Other efflux transporter genes, such as the various multidrug resistant transport genes, *ABCB1, ABCC1, ABCC2* and *ABCG2* have been associated with drug resistance, by increasing in metabolism of doxorubicin. These genes were discussed earlier under the FAC protocol (Section 4.8.3.).

CBR1 and *CBR3* are responsible for the metabolism of anthracyclines. Polymorphisms in these enzymes have been shown to result in lower enzymatic activity and a higher risk of cardiotoxicity. On the other hand, an increase in enzyme activity can reduce the exposure time to the anthracycline, and result in treatment failure, thereby having little to no effect on the reduction of a tumour.[144, 145, 146] The risk alleles for the three CBR gene polymorphisms were found to have a relatively high allele frequency in the study population (rs9024 – G (0.99), rs8133052 – A (0.29), rs1056892 – A (0.48)). Table 20 summarises the variations associated with the metabolism of anthracyclines.

The *UGT1A6*4* rs17863783 - T allele has been marked as a high-risk variant for anthracycline toxicity.[154] The variant has been reported to result in a 20-30% reduction in glucuronidation activity, which may lead to accumulation of reactive oxygen species and toxic alcohol metabolites.[154] The T allele frequency was higher in the study and African population (0.13), compared to the European population (0.02). Testing for this variant in the African population would be essential.

Table 20. Summary of polymorphisms affecting the metabolism of anthracycline and allele frequency of the study cohort of the identified risk allele.

dbSNP	Gene (star allele)	Risk Allele	Freq.	Details
rs1045642	ABCB1	G	0.89	G - associated with an increased metabolism of doxorubicin.[155]
rs714368	SLC22A16	T	0.56	T - associated with a decreased exposure to doxorubicin.[153]
rs9024	CRB1	G	0.99	G - associated with increased clearance of doxorubicin and decreased exposure.[156]
rs8133052	CRB3	G	0.71	GG genotype and breast cancer who are treated with doxorubicin: 1) may have increased metabolism of doxorubicin 2) may have less tumour reduction 3) may have decreased severity of neutropenia as compared to patients with the AA genotype.[157]
rs1056892	CRB3	G	0.52	G - associated with and increased risk of cardiac damage after anthracycline exposure.[158]
rs17863783	UGT1A6(*4)	T	0.13	A - associated with anthracycline associated cardiotoxicity.[154]

NCBI: National Centre for Biotechnology Information

CHAPTER 5: CONCLUSION

The 1000 Genomes Project was a major scientific collaboration which produced a vast amount of information in terms of genetic variability between participating countries. However, Africa and more particularly SSA was highly unrepresented in these studies. Africa, considered to be the cradle of humankind, is expected to contain high levels of genetic diversity and novel variates within the population. The 1000 Genomes Project database (> 2500 genomes) remains the most comprehensive database, even though the southern hemisphere of the African continent is underrepresented.

The 1000 Genomes Project uncovered the presence of many rare pharmacogenetic variants that are not found in European populations. This highlighted the need for further investigations into the African and SSA populations. The literature review conducted by Rajman *et al.* investigated the genetic diversity within the African population with a particular focus on cytochrome P450.[59] Of the 74 papers cited, only eight involved African or African-American populations, and of these one mentioned the Xhosa and their distinct clustering outside of the African, Asian and Caucasian populations. One of the objectives of the study was to assess the extent of genetic variation between the SSA population and the population of European ancestry, as reported in the 1000 Genomes Project. A recent study by de Rocha *et al.* (referred to has the High Coverage African Dataset or HAAD), mined the data of 458 high coverage whole genome sequences to investigate the impact of variations in ADME genes in the SSA population, and is the first study characterizing the landscape of PGx genes.[98] The SSA population was represented by South Africa (n = 157), Namibia/Botswana (n = 53) and Zambia (n = 11). This study highlighted that populations samples from Southern Africa had the greatest number of novel single nucleotide variants identified, embedding the need to investigate this population group more extensively.

This study confirmed that 88% of the SNPs selected for the study showed a significant difference ($P < 0.05$) between the study and European populations, reflecting inter-population variability, which is in agreement with previous reports. Furthermore, the presence of intra-population genetic diversity for the population of African ancestry

and the study population was confirmed, with 48% of the SNPs showing a significant difference ($P < 0.05$). Intra-population diversity was also highlighted, through the allele frequencies of minor alleles when dividing the study population into linguistic groups which are geographically separated. This further speaks to the genetic diversity within the African population. The HAAD study, highlighted clinical actionable variants with PharmGKB evidence levels 1A/1B. Eleven of the ADME variants listed in the HAAD study overlapped with the selected SNPs in the present study; rs35742686 (*CYP2D6*3*), rs3892097 (*CYP2D6*4*), rs1065852 (*CYP2D6*10*), rs28371706 (*CYP2D6*17*), rs4244285 (*CYP2C19*2*), rs12248560 (*CYP2C19*17*), rs776746 (*CYP3A5*3*), rs3745274 (*CYP2B6*6*), rs3918290 (*DPYD*), rs115232898 (*DPYD*) and rs1142345 (*TPMT*). Our observed minor allele frequencies were similar to a large extent except for rs3918290 (*DPYD*), where the minor allele was not found in the present study, but in the HAAD study this allele was present at low frequency (0.0011). The HAAD study observed intra-population variation between the four African population samples namely; Southern Africa, South/Central Africa, Far West Africa and West Africa.[98]

The present study highlighted the difficulties of target-based genotyping, particularly in relation to SNPs that show a high degree of homology and presence of pseudogenes such as the *CYP2D6* gene. Two of the *CYP2D6* polymorphisms, *CYP2D6*10* and *CYP2D6*17,* which are known to be tri- and quad-allelic SNPs causes genotyping uncertainty. Tri-allelic SNPs are probably more common than previously recognised, thus pharmacogenetic studies focused on known polymorphisms may be less reliable. Some data may have potentially been missed given the genetic variability across Africa. This includes unidentified variations unique to the African subpopulations. The HAAD study noted high impact variants that were found to be rare or ultra-rare that were in agreement with our study cohort, namely: *DPYD* (rs3918290) and *TPMT* (rs1142345). The choice of high impact variants may have differed from our selection, as other rare variants found in the present study include; *CYP2C9* (rs79000194), *CYP2C8* (rs10509681) and *CYP2B6* (rs12721655).

Custom SNP designs also pose a challenge, for analysis, as they have not been tested or validated prior to being incorporated on the array. Analysis of these SNPs could thus be challenging, even though quality control checks are put in place. *SULT1A1*1*

encodes for the sulfotransferase 1A1, which catalyses the sulfate conjugate of many hormones, neurotransmitters, drugs and xenobiotic compounds, and forms part of phase II detoxification. This gene was found to show 100% homology to the *STP1* gene. The Genome Aggregation Database (gnomAD) aggregates and harmonizes both exome and genome sequencing data from a variety of large-scale sequencing projects.[159] According to gnomAD, the C allele of *SULT1A1*1* (rs1042028), is present in 83% of the African population, with the T allele having a prevalence of 17%. In this study population the prevalence of the C allele was 77%, and the T allele 23%. Unfortunately, no conclusion could be drawn from any of these SNPs, due to the SNPs failing the HW test. There is also a certain error frequently associated with data generated using TaqMan methodology, where inaccurate genotype calling could occur if individual genotypes fall between the three main genotype clusters.[118] Recording genotypes accurately reflects the true genotypes, however, currently all methods for DNA sequencing are prone to error of this type, resulting in procedural bias.[160] SNPs selected for customized arrays, may also lead to SNP ascertainment bias, and may cause a shift toward intermediate frequency allele, however this can be overcome with large sample sizes. Genome sequencing studies have indicated the difficulty in predicting allele frequencies in admixed populations, such as the SSA population.[119]

Array-based technologies for targeted genotyping have several limitations when being applied as high throughput methods in precision medicine applications. Custom designed arrays would be more beneficial than commercialized arrays, as many were not designed with the African population in mind. Recently the H3Africa consortium in collaboration with Illumina, launched the Infinium™ H3Africa Consortium Array v2. This array allows for the identification of genetic associations with common and rare traits among African populations and has been generated from sequencing data from the 1000 Genomes Project, H3Africa and Wellcome Sanger Institution. Applied Biosystems, also launched an Axiom Pan-African cancer research array, which was developed in collaboration with the Men of African Descent and Carcinoma of the Prostate (MADCaP) consortium. These types of arrays are recommended for genome wide association studies (GWAS). The arrays are not only able to detect SNPs, but also copy number variations, indels and deletions. The purpose of the array was to enable better association of genetic variants with prostate and other cancer in the African populations. The above mentioned collaborative projects are two great

initiatives for high throughput studies, as the lack of novel variant identification remains problematic.

The study further highlighted the individual effect of various polymorphisms on particular anticancer drug classes, the importance thereof relates to the impact that a single gene or group of genes may have on the metabolism of a drug/s. *CYP2D6*, *CYP3A4* and *CYP3A5*, are responsible for the metabolism of over 60% of all drugs and remain essential genetic markers for drug metabolism. The function of the gene and the role it plays in the metabolism of the drug, can alter the effect allele which may be either beneficial or detrimental. For example, *CYP3A4* can be involved in the conversion of a prodrug to an active metabolite, thus the polymorphism would affect the amount of active metabolite present in the system. On the other hand, *CYP3A4* could be responsible for converting an active metabolite to an inactive metabolite for excretion. In this event the polymorphism could be seen to be beneficial, as fast metabolisers would remove the drug from the system quicker, also this could also be detrimental as the drug would have less time to be effective. Albeit in another instances, poor metabolism could be beneficial to allow for an increase in exposure time to the drug. The detrimental effect of the polymorphism, the accumulation of the drug on the individual's system. The effect is thus drug and pathway specific. The effect of transport genes is underestimated, however these genes (SLC transporter genes and multi-drug resistance genes) are the essential regulators of the influx / efflux of molecules into and out of the cell.

The ultimate goal of personalised medicine is to tailor treatment regimens to the individual, in order to improve treatment outcomes. Considering the intra-ethnic genetic diversity observed in the present study, personalised drug-development might not be feasible in SSA. Personalising medicine for each individual based on their genetic makeup will create a bottleneck in the treatment of patients. However, screening for influential polymorphisms prior to treatment, could result in the difference between treatment success or failure. It is also important to understand that a multifactorial approach is required, to achieve a tailored approach to precision medicine, especially for complex diseases such as cancer. The clinical implementation of pharmacogenetics should not solely be based on genetic data. Drug selection should incorporate the histopathology data, for example in the treatment of BrCa,

where treatment options include TAM, platinum drugs or taxanes. Before conducting DNA testing, it is crucial to understand the immunohistochemistry of the tumour in conjunction with gene expression patterns of the tumour using RNA-based microarray. A recent study by Grant *et al.* suggested the reclassification of early stage breast cancer into treatment groups by combining the use of immunohistochemistry and microarray analysis such as the MammaPrint/BluePrint (MP/BP), which measures the functional integrity oestrogen receptors, which can identify patients with false-positive tumours who are resistant to hormone therapy.[161] Once all this information is available, the information can be put into clinical practice.

It is important to note that genomic data is but a subset of the bigger picture. Our bodies are a network, which is not only affected by our genetic characteristics but also external environmental factors. Thus, we cannot consider the genotypic data in isolation. Some SNPs play a more dominant role than others in a drugs pathway. Penetrance refers to the frequency in which a gene is expressed in a population, and is expressed as a percentage of the population that possess the genes and develops a corresponding phenotype. A low penetrance gene may not be expressed even if the trait is dominant. It also may not be expressed when the trait is recessive and the gene responsible for the trait is present on both chromosomes. Penetrance may vary from person to person and its expression may also depend on the person's age. Penetrance will be 100% if all the individuals in the population show the expected phenotype. When it is below 100%, it is referred to as incomplete penetrance which may be as a result of modifiers, epistatic genes, suppressors or environmental factors. A complete omics approach is required, with future studies focusing on epigenetic markers, which can be achieved through bisulphate sequencing. Transcriptomics to assess the expression profiles and transcription of mRNA, through RNA sequencing or expression arrays will be useful. Clinical data also needs to be taken into account, in conjunction with morphological, biochemical and physical traits to establish the link between the various factors. Furthermore, clinical studies with large population sizes are required, to ensure reproducibility of data.

CHAPTER 6: STUDY LIMITATIONS AND RECOMMENDATIONS

Despite the important new insights gained through this study, technology limitations and potential bias in the study is acknowledged. This chapter summarises some of the strengths and weakness of the present study and indicates shortcomings with future prospects for pharmacogenomics studies.

6.1   Methodology strengths

The TaqMan OpenArray™, allowed for accurate and reliable genotyping across a large cohort study and for several SNPs simultaneously, at a fraction of the cost. Although loading optimization was required, once the technique was mastered the array allowed for quick sample processing and analysis.

6.2   Methodology limitations

One of the limitation discussed was the introduction of statistical bias that could potential be introduced through custom array designs. However, we need not observe any effect of this, as the align to the intermediate frequency. A total of 18 samples were excluded from the study due to amplification failure being greater than 10% across the sixty SNPs. The exact reason for the absence of amplification could not be determined, as these samples passed the QC criteria and showed successful plate loading. The greatest limitation to the study was the limited number of SNPs that could be investigated. Therefore potential deletions or novel gene variations could not be determined using the current TaqMan OpenArray™ technology. The current technology is also unable to detect copy number variations, indel and large region deletions. New arrays are available, at an elevated cost per sample. The TaqMan OpenArray™ technology has a current cost of approximately R50.00 per sample, where new arrays from Illumina and Thermo Fisher Scientific (although far more comprehensive) are in the region of R1200.00 per sample.

Whole genome sequencing or target gene sequencing of the samples which showed poor amplification, particularly across the CYP P450 genes, would enable the

identification of potential novel variants or gene deletions. With every study, there is always the aspect of bias that could be introduced. Design bias or SNP ascertainment bias, through the selection of candidate genes, is not considered problematic statistically in genome-wide associated studies. Candidate genes selection was based on literature review from international published data, on American, European and Asian cohorts, due to the limited genetic data available on the African population. This study may therefore prove or disprove observations seen elsewhere. Sample size and statistical bias were prevented through a large population size, and the HWE test, to prevent inaccurate conclusions. However, the data obtained in this study could pave the way for further funding opportunities to conduct larger genome-wide associated studies, and to assess social-environmental parameters.

6.3   Study cohort

The greatest strength of the study lies in the size of the cohort. This is the first study to the researcher's knowledge that investigated the prevalence of several pharmacogenes in the South African population containing more than 700 samples. The use of secondary data, especially during the Covid-19 pandemic, presented very little delays in completing the study within the stipulated period. The one disadvantage of using secondary data lies in the accuracy of data collection. Samples were from four different studies, collected by various individuals. When participants where asked about their home languages, the classification of "Bantu speaking" was sufficient for one study, where for the present study it would have been more beneficial to know the exact mother tongue. This would have strengthened the intrapopulation analysis, as 24% of the study cohort was excluded from this analysis as a result of incomplete data.

6.4   Recommendation for future research

Several published articles have eluded to future focus of ADME studies and the way African genetic diversity can be explored, below five major areas are highlight that future research should be centred on, in order to grow the SSA genetic database and understanding in pharmacogenomics.[9, 59, 94] Firstly, large scale sequencing studies are required with a broader geographical cohort. The majority of pharmacogenomic studies to date have either analysed individual candidate genes or a subset of genetic

variants using genotyping assays to reduce cost and allow for a high sample throughput. Sequencing technologies allow an assessment of the full spectrum of variation present in a given population and enables the detection of rare genetic variation on genes. Although sequencing technologies are the gold standard, they are still very costly. Africa and SSA are resource scarce and access to these newer technologies are quite limited. The cost of whole human genome sequencing in Africa is between R13 000.00 and R45 000.00 depending on the technology used. Illumina in America and the Beijing Genomics Institute in China are able to offer WGS at a cost of $650 per sample (approximately R9 750.00), which is far less than current facilities in Africa are able to offer sequencing at. However, obtaining permits to send samples overseas for sequencing is also costly, and the COVID-19 pandemic has resulted in lengthy delays in obtaining export permits. Whole exome sequencing (WES) is a cheaper alternatively to WGS, however only variates in the coding region of the genome can be assessed. WES is unable to detect genetic variation in the non-coding regions of the genome, and it is also limited by the type of mutation it is able to detect. Currently WGS sequencing strategies employ paired end sequencing of either 100 or 150 bps. Sole reliance on massively parallel, short-read sequencing may, however, prove insufficient despite the optimization of scientific protocols and the technology itself. It will be important to utilize long-read sequencing methods such as Pacific Biosystem and MGI, which offer read length in excess of 150 bps.

Secondly, it is important to grow capacity not only in South Africa, but in Africa as well, through training and the establishment of facilities that are able to conduct pharmacogenetic studies. Community engagement is also important, building trust to help potential study participants understand the research taking place and the role they play. Collaboration between healthcare professionals and researchers would be required, including doctors, nurses, social workers and genetic counsellors. Ideally, and a goal that many local researchers share is to keep African genomic data within Africa and to encourage local researchers to utilize local centres for sequencing. However, there is a general idea that local facilities do not have the expertise to conduct world class sequencing data. This is a great misperception that needs to be curbed. The Western Cape alone is home to three genomic research centers, with scientist receiving both training aboard and locally on next generation sequencing

techniques. This also goes hand in hand with resource development, such as a nation database to store and share sequencing data similar to the 1000 Genomes project.

Thirdly, pharmacovigilance, is another field of science that is highly under practiced on the African continent. It plays a vital role in the healthcare system involving the assessment, monitoring and discovery of drug interactions. It also aids in the discovery of new SNPs affecting drug metabolisms which are not routinely identified. Both researchers and clinicians should explore existing data bases, and report new ADRs in these databases, such as VigiBase®. An extended registry of unexplained adverse events in African patients would be very valuable as it could allow for further pharmacogenetic investigations to be undertaken, greatly facilitating the identification of clinically relevant events that could have pharmacogenetic underpinnings.

Fourthly, clinical based studies are required to establish if or how and to what extent treatment outcome are affected by genetic variations detected in the SSA. Clinical trials should be extended to include participants of African ancestry. The difference in frequency and distribution of some alleles may also influence the outcome of clinical trials and consequently the clinical development of drugs in Africa. This should also be followed with metabolomic studies, which can be aligned with genomic and phenotypic data. Point four and five can function jointly in order to build a database for clinical pharmacogenetic guidelines that can potentially be implemented for the African or SSA population.

Finally, the pharmaceutical industry needs to become more involved in raising awareness of the genetic diversity in Africa and support local clinicians and researchers. Establishment of a global consortium of pharmaceutical companies would be beneficial, to engage with local (African) and global regulatory authorities, as well as other relevant government groups. The main aim of these types of consortiums would be to share ideas and resources of expert academics and clinical centres to develop recommendations for the practical use of pharmacogenetic data in order to optimize the treatment of patients across Africa. Knowledge of the prevalence of known variants of clinical relevance in conjunction with our knowledge of the prevalence of known variants of clinical relevance will aid in guiding new policy developments for drug selection and dosing in African populations on the basis of

pharmacogenetic principles and strategies aimed at improving drug safety and efficacy, creating a sustainable governance of pharmacogenomics in Africa.

6.5   Principles in practice

Healthcare professionals, government and pharmaceutical companies need to work together holistically. Local governments should work alongside local and international pharmaceutical companies (such as Pfizer, Cipla, Aspen, Sanofi and Johnson & Johnson), to draft and implement policies that will enforce the reporting of ADRs in terms of pharmacovigilance requirements. Healthcare professionals need to be aligned with policies that enforce reporting of ADRs and obtain the relevant training. Polypharmacy is another area that is significantly understudied and should be included in pharmacovigilance reporting. Polypharmacy should not only focus on drug-drug interaction, but also drug-food interaction especially considering the extensive use of natural medicine. As investment companies into the economy of Africa, pharmaceutical companies should source local researcher institutions to conduct the necessary research into the cause and effect of reported ADRs. As a resource scare continent, investing into research through funding opportunities would enable and encourage local capacity building.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. A Cancer Journal for Clinicians. 2020;70(1):7-30.
2. Gopal S, Sharpless NE. Cancer as a global health priority. Journal of the American Medical Association. 2021.
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. A Cancer Journal for Clinicians. 2021;71(1):7-33.
4. Crona D, Innocenti F. Can knowledge of germline markers of toxicity optimize dosing and efficacy of cancer therapy? Biomarkers in Medicine. 2012;6(3):349-362.
5. Ampadu HH, Hoekman J, de Bruin ML, Pal SN, Olsson S, Sartori D, Leufkens HG, Dodoo AN. Adverse drug reaction reporting in africa and a comparison of individual case safety report characteristics between africa and the rest of the world: Analyses of spontaneous reports in vigibase®. Drug Safety. 2016;39(4):335-45.
6. Wiffen P, Gill M, Edwards J, Moore A. Adverse drug reactions in hospital patients. A systematic review of the prospective and retrospective studies. Bandolier Extra. 2002:1-16.
7. Bogolubova S, Padayachee N, Schellack N. Knowledge, attitudes and practices of nurses and pharmacists towards adverse drug reaction reporting in the south african private hospital sector. Health SA Gesondheid. 2018;23(0):a1064.
8. Mini E, Nobili S. Pharmacogenetics and personalised medicine: Maintain a critical approach. Prescrire Int. 2013;22(139):161-3.
9. Tshabalala S, Choudhury A, Beeton-Kempen N, Martinson N, Ramsay M, Mancama D. Targeted ultra-deep sequencing of a south african bantu-speaking cohort to comprehensively map and characterize common and novel variants in 65 pharmacologically-related genes. Pharmacogenetics Genomics. 2019;29(7):167-178.
10. Aminkeng F, Ross CJ, Rassekh SR, Brunham LR, Sistonen J, Dube MP, Ibrahim M, Nyambo TB, Omar SA, Froment A, Bodo JM, Tishkoff S, Carleton BC, Hayden MR. Higher frequency of genetic variants conferring increased risk for adrs for commonly used drugs treating cancer, aids and tuberculosis in persons of african descent. Pharmacogenomics Journal. 2014;14(2):160-70.
11. Word Health Organization. Cancer [fact sheet]. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer; 2018.
12. Latest gobal cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018 [press release]. International Agency for Research on Cancer2018.
13. Bay F, Ferlay J, Soerjomataram I, Siegel RL, Torren LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortaility worldwide for 36 cancers in 185 countries. A Cancer Journal for Clinicians. 2018;68:394-424.
14. Ferlay J, Colombet M, Sorjomataram I. Global and regional estimates of the incidence and mortality for 38 cancers: Globocan 2018. Lyon: International Agency for Research on Cancer/World Health Organization; 2018.
15. Herbst MC. Fact sheet on the top ten cancers per population group. Cancer Association of South Africa (CANSA): The National Cancer Registry, 2014; 2018.
16. The National Cancer Registry. Cancer in south africa 2014 full report. National Institute for Communicable Diseases, National Health Laboratory Services; 2014.

17. National health act (act no 61 of 2003): Regulations relating to cancer registration., (2011).
18. Singh E, Sengayi M, Urban M, Babb C, Kellett P, Ruff P. The south african national cancer registry: An update. The Lancet. 2014;15.
19. Foerster M, Anderson BO, McKenzie F, Galukande M, Anele A, Adisa C, Zietsman A, Schuz J, dos Santos Silva I, McCormack V. Inequities in breast cancer treatment in sub-saharan africa: Findings from a prospective multi-country observational study. Breast Cancer Research. 2019;21(1):93.
20. Union for Internation Cancer Control. World cancer declaration - progress report 2016. 2016.
21. Goldstein DA, Clark J, Tu Y, Zhang J, Fang F, Goldstein R, Stemmer SM, Rosenbaum E. A global comparison of the cost of patented cancer drugs in relation to global differences in wealth. Oncotarget. 2017;8(42):71548-71555.
22. The American CancerSociety. About prostate cancer: What is prostate cancer? : American Cancer Society; 2016 [cited 11 June 2019]. Available from: https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html.
23. Patel JN, Wiebe LA, Dunnenberger HM, McLeod HL. Value of supportive care pharmacogenomics in oncology practice. Oncologist. 2018;23(8):956-964.
24. Fact sheet on prostate cancer [press release]. Cancer Association of South Africa (CANSA)2018.
25. The American Cancer Society. Understanding a breast cancer diagnosis: Types of breast cancedr 2017 [cited 11 June 2019]. Available from: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/types-of-breast-cancer.html#written_by.
26. Fact sheet on breast cancer in women [press release]. Cancer Association of South Africa (CANSA) 2018.
27. Ameen S, Qadir MI, Ahmad B. Pharmacogenomic approaches in the treatment of breast cancer by tamoxifen. Pak J Pharm Sci. 2012;25(2):469-76; André F, Ciccolini J, Spano JP, Penault-Llorca F, Mounier N, Freyer G, Blay JY, Milano G. Personalized medicine in oncology: Where have we come from and where are we going? Pharmacogenomics. 2013;14(8):931-9.
28. Carels N, Spinassé LB, Tilli TM, Tuszynski JA. Toward precision medicine of breast cancer. Theoretical Biology and Medical Model. 2016;13:7.
29. Maughan KL, Lutterbie MA, Ham PS. Treatment of breast cancer. American Family Physician. 2010;81(11):1339-1346.
30. Kivisto KT, Kroemer HK, Eichelbaum M. The role of human cytochrome p450 enzymes in the metabolism of anticancer agents: Implications for drug interactions. British Journal of Clinical Pharmacology. 1995;40:523-530.
31. Ahmed S, Zhou Z, Zhou J, Chen S-Q. Pharmacogenomics of drug metabolizing enzymes and transporters: Relevance to precision medicine. Genomics Proteomics Bioinformatics. 2016;14:298-313.
32. Salim Jabir R, Naidu R, Ahmad Annuar MAB, Ho GF. Pharmacogenetics of taxanes: Impact of gene polymorphisms of drug transporters on pharmacokinetics and toxicity. Pharmacogenomics. 2012;13(6):1979-1988.
33. Housman G, Byler S, Heerboth S, Lapinska K, Longacre M, Snyder N, Sarkar S. Drug resistance in cancer: An overview. Cancers. 2014;6:1769-1972.
34. Ruwali M. Pharmacogenetics and cancer treatment: Progess and prospects. In: Nalbantoglu DS, editor. Molecular medicine 2019.

35. Cyprotex. Adme guide 2018 [cited 23 July 2019]. Available from: https://www.cyprotex.com/guides.
36. Campbell JE, Cohall D. Pharmacodynamics—a pharmacognosy perspective. Press A, editor2017.
37. Waller DG, Sampson AP. Medical pharmacology and therapeutics. Edition F, editor: Elsevier; 2018.
38. Collins FS, Varmus H. A new initiative on precision medicine. The New England Journal of Medicine. 2015;372(9):793-795.
39. Staa T, Kaplan, editors. Priority medicines for europe and the world: 2013 update 2013.
40. Jameson LJ, Longo DL. Precision medicine — personalized, problematic, and promising. The New England Journal of Medicine. 2015;372(23):2229-2234.
41. Precision medicine in cancer treatment: National Cancer Institute; 2017 [cited 10 April 2019]. Available from: https://www.cancer.gov/about-cancer/treatment/types/precision-medicine.
42. Grissinger M. The five rights - a destination without a map. Pharmacy and Therapeutics. 2010;35(10):542.
43. Calvo E, Walko C, Dees EC, Valenzuela B. Pharmacogenomics, pharmacokinetics, and pharmacodynamics in the era of targeted therapies. Am Soc Clin Oncol Educ Book. 2016;35:e175-84.
44. Hughes KS, Cusack JC. Genetics, genomics, and pharmacogenomics. Annuals of Surgical Oncology. 2015;22(11):3414-7.
45. Stearns V, Davidson NE, Flockhart D. Pharmacogenetics in the treatment of breast cancer. The Pharmacogenomics Journal. 2004;4:143-153.
46. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, Peterson JF, Van Driest SL. Pharmacogenomics. Genomic Medicine. 2019;2.
47. Vucic EA, Thu KL, Robison K, Rybaczyk LA, Chari R, Alvarez CE, Lam WL. Translating cancer 'omics' to improved outcomes. Genome Research. 2012;22(2):188-95.
48. Morganti S, Tarantino P, Ferraro E, D'Amico P, Duso BA, Curigliano G. Next generation sequencing (ngs): A revolutionary technology in pharmacogenomics and personalized medicine in cancer. Adv Exp Med Biol. 2019;1168:9-30.
49. Salem RM, Rodriguez-Murillo L. Human genome project. In: Gellman MD, Turner JR, editors. Encyclopedia of behavioral medicine. New York, NY: Springer New York; 2013. p. 1003-1004.
50. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics. 1999;22:231-238.
51. Duran I, Hagen C, Arranz J, Apellaniz-Ruiz M, Pérez-Valderrama B, Sala N, Lainez N, García-Del Muro X, Noguerón E, Climent M, Maroto P, Font A, García-Donas J, Gallardo E, López-Criado P, González Del Alba A, Sáez MI, Vázquez S, Luque R, Rodríguez-Antona C. Snps associated with activity and toxicity of cabazitaxel in patients with advanced urothelial cell carcinoma. Pharmacogenomics. 2016;17(5):463-71.
52. Mckenna D. Personalized medicine at coriell, DNA, genes, and snps: Coriell Institute for Medical Research; 2016 [cited 21 November 2020]. Available from: https://www.coriell.org/personalized-medicine/dna-genes-and-snps.

53. Miko I. Phenotype variability: Penetrance and expressivity. Nature Education. 2008;1(1):137.
54. Wu AH. Drug metabolizing enzyme activities versus genetic variances for drug clinical pharmacogenomic relevance. Clinical Proteomics. 2011;8(12):1-9.
55. Sim SC, Kacevska M, Ingelman-Sundberg M. Pharmacogenomics of drug-metabolizing enzymes: A recent update on clinical implications and endogenous effects. Pharmacogenomics Journal. 2013;13(1):1-11.
56. Caudle KE, Sangkuhl K, Whirl-Carrillo M, Swen JJ, Haidar CE, Klein TE, Gammal RS, Relling MV, Scott SA, Hertz DL, Guchelaar HJ, Gaedigk A. Standardizing cyp2d6 genotype to phenotype translation: Consensus recommendations from the clinical pharmacogenetics implementation consortium and dutch pharmacogenetics working group. Clinical and Translational Science. 2020;13(1):116-124.
57. Kivistö KT, Kroemer HK, Eichelbaum M. The role of human cytochrome p450 enzymes in the metabolism of anticancer agents: Implications for drug interactions. British Journal of Clinical Pharmacology. 1995;40(6):523-30.
58. Caudle KE, Sangkuhl K, Whirl-Carrillo M, Swen JJ, Haidar CE, Klein TE, Gammal RS, Relling MV, Stuart A. Scott, Daniel L. Hertz, Henk-Jan Guchelaar, Gaedigk A. Standardizing cyp2d6 genotype to phenotype translation: Consensus recommendations from the clinical pharmacogenetics implementation consortium and dutch pharmacogenetics working group. Clinical and Translational Science. 2020;13:116–124.
59. Rajman I, Knapp L, Morgan T, Masimirembwa C. African genetic diversity: Implications for cytochrome p450-mediated drug metabolism and drug development. EBioMedicine. 2017;17:67-74.
60. Garcia-Casado Z, Guerrero-Zotano A, Llombart-Cussac A, Calatrava A, Fernandez-Serra A, Ruiz-Simon A, Gavila J, Climent MA, Almenar S, Cervera-Deval S, Campos J, Albaladejo CV, Llombart-Bosch A, Guillem V, Lopez-Guerrero JA. A polymorphism at the 3'-utr region of the aromatase gene defines a subgroup of postmenopausal breast cancer patients with poor response to neoadjuvant letrozole. BioMed Central Cancer. 2010;10(36):1-11.
61. Simonsson M, Veerla S, Markkula A, Rose C, Ingvar C, Jernstrom H. Cyp1a2 - a novel genetic marker for early aromatase inhibitor response in the treatment of breast cancer patients. BioMed Central Cancer. 2016;16(256):1-12.
62. Yin B, Li Y, Lu P, Pei K, Liang J, Liu H, Liu X, Xiao J. Advances in pharmacogenoic studies of decetaxel-a review. International Journal of Clinical Medicine. 2017;10(5):7450-7456.
63. Tulsyan S, Chaturvedi P, Agarwal G, Lal P, Agrawal S, Mittal RD, Mittal B. Pharmacogenetic influence of gst polymorphisms on anthracycline-based chemotherapy responses and toxicity in breast cancer patients: A multi-analytical approach. Molecular and Diagnostic Therapy. 2013;17(6):371-9.
64. Yin J-Y, Huang Q, Zhao Y-C, Zhou H-H, Liu Z-Q. Meta-analysis on pharmacogenetics of platinum-based chemotherapy in non small cell lung cancer (nsclc) patients. PLoS One. 2012;7(6):1-12.
65. McWhinney SR, Goldberg RM, McLeod HL. Platinum neurotoxicity pharmacogenetics. Molecular Cancer Therapy. 2009;8(1):10-16.
66. Giacomini KM, Yee SW, Mushiroda T, Weinshilboum RM, Ratain MJ, Kubo M. Genome-wide association studies of drug response and toxicity: An opportunity for genome medicine. Nature Reviews Drug Discovery. 2017;16(1):1.

67. Drozda K, Pacanowski MA, Grimstein C, Zineh I. Pharmacogenetic labeling of fda-approved drugs: A regulatory retrospective. JACC. Basic to translational science. 2018;3(4):545-549.

68. Cacabelos R, Cacabelos N, Carril JC. The role of pharmacogenomics in adverse drug reactions. Expert Reviews in Clinical Pharmacology. 2019;12(5):407-442.

69. Bachtiar M, Lee CGL. Genetics of population differences in drug response. Current Genetic Medicine Reports. 2013;1(3):162-170.

70. Abdullah-Koolmees H, van Keulen AM, Nijenhuis M, Deneer VHM. Pharmacogenetics guidelines: Overview and comparison of the dpwg, cpic, cpnds, and rnpgx guidelines. Frontiers in Pharmacology. 2021;11.

71. DrugBank. Abiraterone acetate 2021 [cited 03 July 2021]. Available from: https://go.drugbank.com/drugs/DB05812.

72. Shiota M, Narita S, Akamatsu S, Fujimoto N, Sumiyoshi T, Fujieara M, Uchiumi T, Habuchi T, Ogawa O, Eto M. Association of missense polymorphism in hsd3b1 with outcomes among men with prostate cancer treated with androgene-deprivation therapy or abiraterone. Journal of the American Medical Association. 2019;2(2):1.

73. Qin S, Liu D, Kohli M, Wang L, Vedell PT, Hillman DW, Niu N, Yu J, Richard M. Weinshilboum, Wang L. Tspyl family regulates cyp17a1 and cyp3a4expression: Potential mechanism contributingto abiraterone response in metastaticcastration-resistant prostate cancer. Clinical Pharmacology & Therapeutics. 2018;104(1):201-210.

74. Reis SS, Carvalho AS, Fernandes R. Pharmacogenomics, cyp2d6, and tamoxifen: A survey of the reasons sustaining european clinical practice paradigms. Medicina (Kaunas). 2019;55(7).

75. DrugBank. Tamoxifen 2021 [cited 03 July 2021]. Available from: https://go.drugbank.com/drugs/DB00675.

76. ARomero-Lorca A, Novillo A, Gaibar M, Bandres F, Fernandez-Santander A. Impacts of the glucuronidase genotypes ugt1a4, ugt2b7, ugt2b15 and ugt2b17 on tamoxifen metabolism in breast cancer patients. PLoS One. 2015;10(7):1-16.

77. Dickschen K, Willmann S, Thelen K, Lippert J, Hempel G, Eissing T. Physiologically based pharmacokinetic modeling of tamoxifen and its metabolites in women of different cyp2d6 phenotypes provides new insight into the tamoxifen mass balance. Frontiers in Pharmacology. 2012;3:92.

78. Artigalás O, Vanni T, Hutz MH, Ashton-Prolla P, Schwartz IV. Influence of cyp19a1 polymorphisms on the treatment of breast cancer with aromatase inhibitors: A systematic review and meta-analysis. BMC Medicine. 2015;13:139.

79. Irahara N, Miyoshi Y, Taguchi T, Tamaki Y, S N. Quantitative analysis of aromatase mrna expression derived from various promoters (i.4, i.3, pii ans i.7) and its association with expression of tnf-alpha, il 6 and cox-2 mrnas in human breast cancer. International Journal of Cancer. 2006;118(8):1915-1921.

80. Ferraldeschi R, Arnedos M, Hadfield KD, A'Hern R, Drury S, Wardley A, Howell A, Evans DG, Roberts SA, Smith I, Newman WG, Dowsett M. Polymorphisms of cyp19a1 and response to aromatase inhibitors in metastatic breast cancer patients. Breast Cancer Research Treatment. 2012;133(3):1191-8.

81. Glaser RL, York AE, Dimitrakakis C. Subcutaneous testosterone-letrozole therapy before and concurrent with neoadjuvant breast chemotherapy: Clinical response and therapeutic implications. Menopause. 2017;24(7):859-864.

82. Dunning AM, Dowsett M, Healey CS, Tee L, Luben RN, Folkerd E, Novik KL, Keleman L, Ogata S, Phoroah P. Polymorhisms associated with circulating sex

hormone levels in postmenopausal women. Journal of the National Cancer Institute. 2004;96(12):936-945; Ji-Rong Long, Nobuhiko Kataoka, Xiao-Ou Shu, Wanqing Wen, Yu-Tang Gao, Quiyin Cai, Zheng W. Genetic polymorhisms of the cyp19a1 gene and breast cancer survival. Cancer Epidemiology Biomarkers and Prevention. 2006;15(11):2115-2122.

83. Precht JC, Schroth W, Klein K, Brauch H, Krynetskiy E, Schwab M, Murdter TE. The letrozole phase 1 metabolite carbinol as a novel probe drug for ugt2b7. Drug Metabolism and Disposition. 2013;41(11):1906-1913.

84. Roco A, Cayun J, Contreras S, Stojanova J, Quinones L. Can pharmacogenetics explain efficacy and safety of cisplatin pharmacotherapy? Frontiers in Genetics. 2014;5(39):1-14.

85. Hoffmeyer S, Burk O, von Richter O, Arnold HP, Brockmoller J, Johne A, Cascorbi I, Gerloff T, Roots I, Eichelbaum M, Brinkmann U. Functional polymorphisms of the human mutlidrug-resistance gene: Multiple sequence variations and correlation of one allele with p-glycoprotein expression and activity *in vivo*. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(7):3473-3478.

86. Whirl-Carrillo M, McDonagh E M, Hebert J M, Gong L, Sangkuhl K, Thorn C F, Altman BR, E. KT. Pharmacogenomics knowledge for personalized medicine. Clinical pharmacology and therapeutics. 2012.

87. Marsh S, McLeod H, Dolan E, Shukla SJ, Rabik CA, Gong L, Hernandez-Boussard T, Lou XJ, Klein TE, Altman BR. Platinum pathway. Pharmacogenetics and genomics. 2009.

88. Thorn CF, Oshiro C, Marsh S, Hernandez-Boussard Tina, McLeod Howard, Klein Teri E, Altman BR. Doxorubicin pathways: Pharmacodynamics and adverse effects. Pharmacogenetics and genomics. 2011.

89. Krens SD, McLeod HL, Hertz DL. Pharmacogenetics, enzyme probes and therapeutic drug monitoring as potential tolls for individualizing taxane therapy. Pharmacogenomics. 2013;14(5):555-574.

90. Varughese LA, Lau-Min KS, Cambareri C, Damjanov N, Massa R, Reddy N, Oyer R, Teitelbaum U, Tuteja S. Dpyd and ugt1a1 pharmacogenetic testing in patients with gastrointestinal malignancies: An overview of the evidence and considerations for clinical implementation. Pharmacotherapy. 2020;40(11):1108-1129.

91. Chan HT, Chin YM, Low S-K. The roles of common variation and somatic mutation in cancer pharmacogenomics. Oncology and Therapy. 2019;7(1):1-32.

92. Thorn CF, Marsh S, Whirl- Carrillo M, McLeod HL, Klein Teri E, Altman BR. Pharmgkb summary: Fluoropyrimidine pathways. Pharmacogenetics and genomics. 2011.

93. Mikkelsen TS, Thorn CF, Yang JJ, Ulrich CM, French D, Zaza G, Dunnenberger HM, Marsh S, McLeod HL, Giacomini K, Becker ML, Gaedigk , Leeder JS, Kager L, Relling MV, Evans W, Klein Teri E, B. AR. Pharmgkb summary: Methotrexate pathway. Pharmcogenetics and genomics. 2011.

94. Radouani F, Zass L, Hamdi Y, Rocha JD, Sallam R, Abdelhak S, Ahmed S, Azzouzi M, Benamri I, Benkahla A, Bouhaouala-Zahar B, Chaouch M, Jmel H, Kefi R, Ksouri A, Kumuthini J, Masilela P, Masimirembwa C, Othman H, Panji S, Romdhane L, Samtal C, Sibira R, Ghedira K, Fadlelmola F, Kassim SK, Mulder N. A review of clinical pharmacogenetics studies in african populations. Pecision Medicine. 2020;17(2):155-170.

95. da Rocha JEB, Lombard Z, Ramsay M. Potential impact of dpyd variation on fluoropyrimidine drug response in sub-saharan african populations. Frontiers in Genetics. 2021;12(162).

96. Rajman I, Knapp L, Hanna I. Genetics diversity in drug transporters: Impact in african populations. Clinical and Translational Science. 2020;13(5):848-860.

97. Ndadza A, Muyambo S, Mntla P, Wonkam A, Chimusa E, Kengne AP, Ntsekhe M, C. D. Profiling of warfarin pharmacokinetics-associated genetic variants: Black africans portray unique genetic markers important for an african specific warfarin pharmacogenetics-dosing algorithm. . Journal of Thrombosis and Haemostasis. 2021;19(12):2957-2973.

98. da Rocha JEB, Othman H, Botha G, Cottino L, Twesigomwe D, Ahmed S, Drögemöller BI, Fadlelmola FM, Machanick P, Mbiyavanga M, Panji S, Wright GEB, Adebamowo C, Matshaba M, Ramsay M, Simo G, Simuunza MC, Tiemessen CT, Baldwin S, Chiano M, Cox C, Gross AS, Thomas P, Gamo F-J, Hazelhurst S. The extent and impact of variation in adme genes in sub-saharan african populations. Frontiers in Pharmacology. 2021;12(366).

99. Joerger M, Huitema ADR, Richel DJ, Dittrich C, Pavlidis N, Briasoulis E, Vermorken JB, Strocchi E, Martoni A, Sorio R, Sleeboom HP, Izquierdo MA, Jodrell DI, Féty R, de Bruijn E, Hempel G, Karlsson M, Tranchand B, Schrijvers AHGJ, Twelves C, Beijnen JH, Schellens JHM. Population pharmacokinetics and pharmacodynamics of doxorubicin and cyclophosphamide in breast cancer patients. Clinical Pharmacokinetics. 2007;46(12):1051-1068.

100. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. Jama. 2013;310(20):2191-4.

101. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Garcia Giron C, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Gonzalez Martinez J, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam Ahamed I A, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P. Ensembl 2021. Nucleic Acids Res. 2020;49(D1):D884-D891.

102. Dbsnp Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 [cited 21 June 2019]. Available from: https://www.ncbi.nlm.nih.gov/gene/.

103. Johns MB, Paulus-Thomas JE. Purification of human genomic DNA from whole blood using sodium perchlorate in place of phenol. Analytical Biochemistry. 1989;180(2):276-278.

104. Nicklas JA, Buel E. Quantification of DNA in forensic samples. Analytical and Bioanalytical Chemistry. 2003;376(8):1160-7.

105. Invitrogen. Qubit(tm) 4.0 fluorometer - user guide. Life Technologies: USA; 2017.

106. Invitrogen. Qubit(r) 3.0 fluorometer - user guide. USA.: Life Technologies; 2014 [cited April 12 2019]. Available from: http://tools.thermofisher.com/content/sfs/manuals/qubit_3_fluorometer_man.pdf .

107. Martins FTA, Ramos PZ, Svidnicki MCCM, Castilho AM, Sartorato EL. Optimization of simultaneous screening of the main mutations involved in non-syndromic deafness using the taqman® openarray™ genotyping platform. BMC Medical Genetics. 2013;14(112).

108. Corporation LT. Taqman openarray genotyping user manual. In: Biosystems A, editor. Part Number 4377476 Rev. E2010. p. 1-158.

109. Panagiotara A, Skoufas E, Chalikiopoulou C, Tsermpini E-E, Bartsakoulia M, Katsila T, Patrinos GP. 15 - implementation of genomic medicine: An international perspective. In: Pyeritz RE, Korf BR, Grody WW, editors. Emery and rimoin's principles and practice of medical genetics and genomics (seventh edition): Academic Press; 2019. p. 369-380.

110. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Wang J, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Lander ES, Altshuler DM, Gabriel SB, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Flicek P, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Bentley DR, Grocock R, Humphray S, James T, Kingsbury Z, Lehrach H, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo M-L, Mardis ER, Wilson RK, Fulton L, Fulton R, Sherry ST, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, McVean GA, Durbin RM, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Schmidt JP, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Auton A, Campbell CL, Kong Y, Marcketta A, Gibbs RA, Yu F, Antunes L, Bainbridge M, Muzny D, Sabo A, Huang Z, Wang J, Coin LJM, Fang L, Guo X, Jin X, Li G, Li Q, Li Y, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Alkan C, Dal E, Kahveci F, Marth GT, Garrison EP, Kural D, Lee W-P, Fung Leong W, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Altshuler DM, Banks E, Bhatia G, del Angel G, Gabriel SB, Genovese G, Gupta N, Li H, Kashin S, Lander ES, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Clark AG, Gottipati S, Keinan A, Rodriguez-Flores JL, Korbel JO, Rausch T, Fritz MH, Stütz AM, Flicek P, Beal K, Clarke L, Datta A, Herrero J, McLaren WM, Ritchie GRS, Smith RE, Zerbino D, Zheng-Bradley X, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball EV, Stenson PD, Bentley DR, Barnes B, Bauer M, Keira Cheetham R, Cox A, Eberle M, Humphray S, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Sudbrak R, Amstislavskiy VS, Herwig R, Mardis ER, Ding L, Koboldt DC, Larson D, Ye K, Gravel S, The Genomes Project C, Corresponding a, Steering c, Production g, Baylor College of M, Shenzhen BGI, Broad Institute of MIT, Harvard, Coriell Institute for Medical R, European Molecular Biology Laboratory EBI, Illumina,

Max Planck Institute for Molecular G, McDonnell Genome Institute at Washington U, Health USNIo, University of O, Wellcome Trust Sanger I, Analysis g, Affymetrix, Albert Einstein College of M, Bilkent U, Boston C, Cold Spring Harbor L, Cornell U, European Molecular Biology L, Harvard U, Human Gene Mutation D, Icahn School of Medicine at Mount S, Louisiana State U, Massachusetts General H, McGill U, National Eye Institute NIH. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

111. Halliburton R. Introduction to population genetics: Pearson/Prentice Hall; 2004.

112. Stats South Africa. Census 2011 statistical release - p0301.4. Pretoria: Stats SA Library Cataloguing-in-Publication (CIP) Data; 2012.

113. Offer SM, Lee AM, Mattison LK, Fossum C, Wegner NJ, Diasio RB. A dpyd variant (y186c) in individuals of african ancestry is associated with reduced dpd enzyme activity. Clin Pharmacol Ther. 2013;94(1):158-66; Muhammad Wasif Saif, Adam M Lee, Steven M Offer, Kathleen McConnell, Valerie Relias, Diasio RB. A dpyd variant (y186c) specific to individuals of african descent in a patient with life-threatening 5-fu toxic effects: Potential for an individualized medicine approach. Mayo Clinic Proceedings. 2014;89(1):131-136.

114. Ensembl. Rs7900194 snp Ensembl release 104 - May 20212021 [cited 28 August 2021]. Available from: https://www.ensembl.org/Homo_sapiens/Variation/Explore?r=10:94941809-94942809;v=rs7900194;vdb=variation;vf=168159511.

115. Ensembl. Rs2032582 snp: Ensembl release 104 - May 2021 © EMBL-EBI; 2021 [Available from: https://www.ensembl.org/Homo_sapiens/Variation/Explore?r=7:87530802-87531802;v=rs2032582;vdb=variation;vf=729605947.

116. Hardy GH. Mendelian proportions in a mixed population. Science. 1908;28(706):49-50.

117. Crow JF. Eighty years ago: The beginnings of population genetics. Genetics. 1988;119(3):473-6.

118. Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu C-F. Detection of genotyping errors by hardy–weinberg equilibrium testing. European Journal of Human Genetics. 2004;12(5):395-399.

119. Lachance J, Tishkoff SA. Snp ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays : news and reviews in molecular, cellular and developmental biology. 2013;35(9):780-786.

120. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

121. Tecza K, Pamula-Pilat J, Lanuszewska J, Butkiewicz D, Grzybowska E. Pharmacogenetics of toxicity of 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer patients. Oncotarget. 2018;9(10):9114-9136.

122. Yang Y, Botton MR, Scott ER, Scott SA. Sequencing the cyp2d6 gene: From variant allele discovery to clinical pharmacogenetic testing. Pharmacogenomics. 2017;18(7):673-685.

123. Kotnis A, Kannan S, Sarin R, Mulherkar R. Case–control study and meta-analysis of sult1a1 arg213his polymorphism for gene, ethnicity and environment interaction for cancer risk. British Journal of Cancer. 2008;99(8):1340-1347.

124. Li B, Shi X, Yuan Y, Peng M, Jin H, Qin D. Ercc1 rs11615 polymorphism increases susceptibility to breast cancer: A meta-analysis of 4547 individuals. Bioscience Repository. 2018;38(3):BSR20180440.

125. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. 2005;6(2):95-108.
126. Wang LY, Cui JJ, Liu JY, Guo AX, Zhao ZY, Liu YZ, Wu JC, Li M, Hu CP, Gao Y, Zhou HH, Yin JY. Gene-gene and gene-environment interaction data for platinum-based chemotherapy in non-small cell lung cancer. Scientific Data. 2018;5:180284.
127. Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, Chen R, Sirota M, Kodama K, Hadley D, Butte AJ. Are minor alleles more likely to be risk alleles? BMC Med Genomics. 2018;11(1):3.
128. Robert F, Pelletier J. Exploring the impact of single-nucleotide polymorphisms on translation. Front Genetics. 2018;9(507).
129. Zeng Z, Bromberg Y. Predicting functional effects of synonymous variants: A systematic review and perspectives. Front Genetics. 2019;10:914-914.
130. Crews KR, Monte AA, Huddart R, Caudle KE, Kharasch ED, Gaedigk A, Dunnenberger HM, Leeder JS, Callaghan JT, Samer CF, Klein TE, Haidar CE, Van Driest SL, Ruano G, Sangkuhl K, Cavallari LH, Müller DJ, Prows CA, Nagy M, Somogyi AA, Skaar TC. Clinical pharmacogenetics implementation consortium guideline for cyp2d6, oprm1, and comt genotypes and select opioid therapy. Clinical Pharmacology Therapy. 2021.
131. Hoffman JM, Dunnenberger HM, Hicks KJ, Caudle KE, Whirl-Carrillo M, Freimuth RR, Williams MS, Klein TE, Peterson JF. Developing knowledge resources to support precision medicine: Principles from the clinical pharmacogenetics implementation consortium (cpic). Journal of Americal Medical Informatics Association. 2016;23(4):796-801.
132. Thorn CF, Klein TE, Altman RB. Pharmgkb: The pharmacogenomics knowledge base. Methods in molecular biology (Clifton, N.J.). 2013;1015:311-320.
133. Binder M, Zhang BY, Hillman DW, Kohli R, Kohli T, Lee A, Kohli M. Common genetic variation in cyp17a1 and response to abiraterone acetate in patients with metastatic castration-resistant prostate cancer. International Journal of Molecular Science. 2016;17(7).
134. Elahe A Mostaghel, Brett T Marck, Stephen R Plymate, Robert L Vessella, Stephen Balk, Alvin M Matsumoto, Peter S Nelson, Montgomery BR. Resistance to cyp17a1 inhibition with abiraterone in castration-resistant prostate cancer: Induction of steroidogenesis and androgen receptor splice variants. Clinical cancer research. 2011;17(18):5913-25.
135. Kim HS, Kim MK, Chung HH, Kim JW, Park NH, Song YS, Kang SB. Genetic polymorphisms affecting clinical outcomes in epithelial ovarian cancer patients treated with taxanes and platinum compounds: A korean population-based study. Gynecology Oncology. 2009;113(2):264-9.
136. Kalikaki A, Kanaki M, Vassalou H, Souglakos J, Voutsina A, Georgoulias V, Mavroudis D. DNA repair gene polymorphisms predict favorable clinical outcome in advanced non-small-cell lung cancer. Clinical Lung Cancer. 2009;10(2):118-23.
137. Huang MY, Huang ML, Chen MJ, Lu CY, Chen CF, Tsai PC, Chuang SC, Hou MF, Lin SR, Wang JY. Multiple genetic polymorphisms in the prediction of clinical outcome of metastatic colorectal cancer patients treated with first-line folfox-4 chemotherapy. Pharmacogenetic Genomics. 2011;21(1):18-25.
138. Mohelnikova-Duchonova B, Melichar B, Soucek P. Folfox/folfiri pharmacogenetics: The call for a personalized approach in colorectal cancer therapy. World Journal of Gastroenterology. 2014;20(30):10316-30.

139. Björn N, Sigurgeirsson B, Svedberg A, Pradhananga S, Brandén E, Koyi H, Lewensohn R, de Petris L, Apellániz-Ruiz M, Rodríguez-Antona C, Lundeberg J, Gréen H. Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia.Pharmacogenomics Journal. 2020;20(2):179-191.

140. Khrunin AV, Moisseev A, Gorbunova V, Limborska S. Genetic polymorphisms and the efficacy and toxicity of cisplatin-based chemotherapy in ovarian cancer patients. Pharmacogenomics J. 2010;10(1):54-61.

141. Park HS, Lim SM, Shin HJ, Cho A, Shin JG, Lee MG, Kim HR, Kim JH, Cho BC. Pharmacogenetic analysis of advanced non-small-cell lung cancer patients treated with first-line paclitaxel and carboplatin chemotherapy. Pharmacogenet Genomics. 2016;26(3):116-25.

142. Lee A, Moon BI, Kim TH. Brca1/brca2 pathogenic variant breast cancer: Treatment and prevention strategies. Annal of Laboratory Medicine. 2020;40(2):114-121.

143. Drögemöller BI, Monzon JG, Bhavsar AP, Borrie AE, Brooks B, Wright GEB, Liu G, Renouf DJ, Kollmannsberger CK, Bedard PL, Aminkeng F, Amstutz U, Hildebrand CA, Gunaretnam EP, Critchley C, Chen Z, Brunham LR, Hayden MR, Ross CJD, Gelmon KA, Carleton BC. Association between slc16a5 genetic variation and cisplatin-induced ototoxic effects in adult patients with testicular cancer. JAMA Oncology. 2017;3(11):1558-1562.

144. Tecza K, Pamula-Pilat J, Lanuszewska J, Grzybowska E. Pharmacogenetics of fac chemotherapy side effects in breast cancer patients. Hereditary Cancer in Clinical Practice. 2015;13:A10.

145. Tecza K, Pamula-Pilat J, Lanuszewska J, Butkiewicz D, Grzybowska E. Pharmacogenetics of toxicity of 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer patients. Oncotarget. 2018;9(10):9114-9136.

146. Frederiks CN, Lam SW, Guchelaar HJ, Boven E. Genetic polymorphisms and paclitaxel- or docetaxel-induced toxicities: A systematic review. Cancer Treatment Review. 2015;41(10):935-50.

147. Chang H, Rha SY, Jeung HC, Im CK, Ahn JB, Kwon WS, Yoo NC, Roh JK, Chung HC. Association of the abcb1 gene polymorphisms 2677g>t/a and 3435c>t with clinical outcomes of paclitaxel monotherapy in metastatic breast cancer patients. Annals of Oncology. 2009;20(2):272-7.

148. Jabir RS, Naidu R, Annuar MA, Ho GF, Munisamy M, Stanslas J. Pharmacogenetics of taxanes: Impact of gene polymorphisms of drug transporters on pharmacokinetics and toxicity. Pharmacogenomics. 2012;13(16):1979-88.

149. Dai D, Zeldin DC, Blaisdell JA, Chanas B, Coulter SJ, Ghanayem BI, Goldstein JA. Polymorphisms in human cyp2c8 decrease metabolism of the anticancer drug paclitaxel and arachidonic acid. Pharmacogenetics. 2001;11(7):597-607.

150. Kus T, Aktas G, Kalender ME, Demiryurek AT, Ulasli M, Oztuzcu S, Sevinc A, Kul S, Camci C. Polymorphism of cyp3a4 and abcb1 genes increase the risk of neuropathy in breast cancer patients treated with paclitaxel and docetaxel. Oncotargets Ther. 2016;9:5073-5080.

151. Leskelä S, Jara C, Leandro-García LJ, Martínez A, García-Donas J, Hernando S, Hurtado A, Vicario JC, Montero-Conde C, Landa I, López-Jiménez E, Cascón A, Milne RL, Robledo M, Rodríguez-Antona C. Polymorphisms in cytochromes

p450 2c8 and 3a5 are associated with paclitaxel neurotoxicity. Pharmacogenomics Journals. 2011;11(2):121-9.

152. Mbatchi LC, Schmitt A, Thomas F, Cazaubon Y, Robert J, Lumbroso S, Brouillet JP, Pourquier P, Chatelut E, Boyer JC, Evrard A. Polymorphisms in slco1b3 and nr1i2 as genetic determinants of hematotoxicity of carboplatin and paclitaxel combination. Pharmacogenomics. 2015;16(13):1439-50.

153. Bray J, Sludden J, Griffin MJ, Cole M, Verrill M, Jamieson D, Boddy AV. Influence of pharmacogenetics on response and toxicity in breast cancer patients treated with doxorubicin and cyclophosphamide. British Journal of Cancer. 2010;102(6):1003-9.

154. Aminkeng F, Ross CJ, Rassekh SR, Hwang S, Rieder MJ, Bhavsar AP, Smith A, Sanatani S, Gelmon KA, Bernstein D, Hayden MR, Amstutz U, Carleton BC. Recommendations for genetic testing to reduce the incidence of anthracycline-induced cardiotoxicity. British Journal of Clinical Pharmacology. 2016;82(3):683-95.

155. Lal S, Wong ZW, Sandanaraj E, Xiang X, Ang PC, Lee EJ, Chowbay B. Influence of abcb1 and abcg2 polymorphisms on doxorubicin disposition in asian breast cancer patients. Cancer Science. 2008;99(4):816-23.

156. Lal S, Sandanaraj E, Wong ZW, Ang PC, Wong NS, Lee EJ, Chowbay B. Cbr1 and cbr3 pharmacogenetics and their influence on doxorubicin disposition in asian breast cancer patients. Cancer Science. 2008;99(10):2045-54.

157. Fan L, Goh BC, Wong CI, Sukri N, Lim SE, Tan SH, Guo JY, Lim R, Yap HL, Khoo YM, Iau P, Lee HS, Lee SC. Genotype of human carbonyl reductase cbr3 correlates with doxorubicin disposition and toxicity. Pharmacogenetics Genomics. 2008;18(7):621-31.

158. Blanco JG, Sun CL, Landier W, Chen L, Esparza-Duran D, Leisenring W, Mays A, Friedman DL, Ginsberg JP, Hudson MM, Neglia JP, Oeffinger KC, Ritchey AK, Villaluna D, Relling MV, Bhatia S. Anthracycline-related cardiomyopathy after childhood cancer: Role of polymorphisms in carbonyl reductase genes--a report from the children's oncology group. Journal of Clinical Oncology. 2012;30(13):1415-21.

159. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, Watts NA, Solomonson M, O'Donnell-Luria A, Baumann A, Munshi R, Walker M, Whelan CW, Huang Y, Brookings T, Sharpe T, Stone MR, Valkanas E, Fu J, Tiao G, Laricchia KM, Ruano-Rubio V, Stevens C, Gupta N, Cusick C, Margolin L, Alföldi J, Armean IM, Banks E, Bergelson L, Cibulskis K, Collins RL, Connolly KM, Covarrubias M, Cummings B, Daly MJ, Donnelly S, Farjoun Y, Ferriera S, Francioli L, Gabriel S, Gauthier LD, Gentry J, Gupta N, Jeandet T, Kaplan D, Karczewski KJ, Laricchia KM, Llanwarne C, Minikel EV, Munshi R, Neale BM, Novod S, O'Donnell-Luria AH, Petrillo N, Poterba T, Roazen D, Ruano-Rubio V, Saltzman A, Samocha KE, Schleicher M, Seed C, Solomonson M, Soto J, Tiao G, Tibbetts K, Tolonen C, Vittal C, Wade G, Wang A, Wang Q, Ware JS, Watts NA, Weisburd B, Whiffin N, Salinas CAA, Ahmad T, Albert CM, Ardissino D, Atzmon G, Barnard J, Beaugerie L, Benjamin EJ, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Chambers JC, Chan JC, Chasman D, Cho J, Chung MK, Cohen B, Correa A, Dabelea D, Daly MJ, Darbar D, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Erdmann J, Esko T, Färkkilä M, Florez J, Franke A, Getz G, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Haiman C, Hanis C, Harms M, Hiltunen M, Holi MM, Hultman CM, Kallela M, Kaprio J, Kathiresan S, Kim B-J, Kim YJ, Kirov

G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lehtimäki T, Loos RJF, Lubitz SA, Ma RCW, MacArthur DG, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Neale BM, Nilsson PM, O'Donovan MC, Ongur D, Orozco L, Owen MJ, Palmer CNA, Palotie A, Park KS, Pato C, Pulver AE, Rahman N, Remes AM, Rioux JD, Ripatti S, Roden DM, Saleheen D, Salomaa V, Samani NJ, Scharf J, Schunkert H, Shoemaker MB, Sklar P, Soininen H, Sokol H, Spector T, Sullivan PF, Suvisaari J, Tai ES, Teo YY, Tiinamaija T, Tsuang M, Turner D, Tusie-Luna T, Vartiainen E, Vawter MP, Ware JS, Watkins H, Weersma RK, Wessman M, Wilson JG, Xavier RJ, Taylor KD, Lin HJ, Rich SS, Post WS, Chen Y-DI, Rotter JI, Nusbaum C, Philippakis A, Lander E, Gabriel S, Neale BM, Kathiresan S, Daly MJ, Banks E, MacArthur DG, Talkowski ME, Genome Aggregation Database Production T, Genome Aggregation Database C. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444-451.

160. Waples RS. Testing for hardy–weinberg proportions: Have we lost the plot? Journal of Heredity. 2014;106(1):1-19.

161. Grant K, Myburgh E, Murray E, Pienaar F, Kidd M, Wright C, Kotze M. Reclassification of early stage breast cancer into treatment groups by combining the use of immunohistochemistry and microarray analysis. South African Journal of Science. 2019;115.

APPENDICES

Appendix I: Ethical approval from the Research Ethics Committee, Faculty of Health Sciences, University of Pretoria, for the present study.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IORG #: IORG0001762 OMB No. 0990-0279 Approved for use through February 28, 2022 and Expires: 03/04/2023.

22 January 2021

Approval Certificate
Annual Renewal

Ethics Reference No.: 3/2020
Title: Prevalence of genetic polymorphisms associated with anti-cancer drug efficacy and toxicity in the South African population

Dear Ms B Sansom

The **Annual Renewal** as supported by documents received between 2020-11-17 and 2021-01-20 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2021-01-20 as resolved by its quorate meeting.

Please note the following about your ethics approval:
- Renewal of ethics approval is valid for 1 year, subsequent annual renewal will become due on 2022-01-22.
- Please remember to use your protocol number (3/2020) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers
MBChB MMed (Int) MPharmMed PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Appendix II: Ethical approval from the Faculty of Health Sciences Research Ethics Committee, University of Pretoria, for the blood bank donor cohort.

FWA 00002567, Approved dd 22 May 2002 and
Expires 24 Jan 2009
IRB 0000 2235 IORG0001762 Approved dd Jan 2006 and
Expires 21 Nov 2008

University of Pretoria
Faculty of Health Sciences Research Ethics Committee
University of Pretoria
Date: 12 / 09 / 2006

Soutpansberg Road   Private Bag x 385
MRC-Building   Pretoria
Room 2 - 19   0001

Number	:	73/2006
Title	:	Genetic Polymorphisms in the General Population
Investigators	:	Prof S W van der Merwe [Principal Investigator] Department of Hepatology and Gastroenterology; University of Pretoria ; Pretoria
		Prof L van Rensburg [Principal Investigator] Department of Human Genetics; University of Pretoria ; Pretoria
		Dr R Bond [Investigator] Department of Gastroenterology; Pretoria Academic Hospital University of Pretoria; Pretoria
		Mr E Fritz ✉ fritz@up.ac.za, T: 012 319 2060 F: 012 328 3600
Sponsor	:	No
VAT No	:	No
Study Degree	:	No

This Protocol and Informed Consent have been considered by the Faculty of Health Sciences Research Ethics Committee, University of Pretoria on 6/09/2006 and found to be acceptable.

Mr P Behari   B.Proc. KZN; LLM – Unisa; (Lay Member)
*Advocate AG Nienaber   (female)BA(Hons) (Wits); LLB; LLM (UP); Dipl.Datametrics (UNISA)
Prof V.O.L. Karusseit   MBChB; MFGP (SA); M.Med (Chir); FCS (SA): Surgeon
*Prof M Kruger   (female) MB.ChB.(Pret); Mmed.Paed.(Pret); PhDd. (Leuven)
Dr N K Likibi   MB.BCh.; Med.Adviser (Gautang Dept.of Health)
Dr F M Mulaudzi   (female) Department of Nursing,
Mrs E.L. Nombe   (female) B.A. CUR Honours; MSC Nursing – UNISA (Lay Member)
*Snr Sr J. Phatoli   (female) BCur (Et.Al) Senior Nursing-Sister
*Dr L Schoeman   (female) Bpharm, BA Hons (Psy), PhD
*Prof J.R. Snyman   MBChB, M.Pharm.Med; MD: Pharmacologist
*Dr R Sommers   (female) MBChB; M.Med (Int); MPhar.Med;
*Prof TJP Swart   BChD, MSc (Odont), MChD (Oral Path) Senior Specialist; Oral Pathology
*Prof C W van Staden   MBChB; Mmed (Psych); MD; FTCL; UPLM; Dept of Psychiatry
*Dr AP van der Walt   BChD, DGA (Pret) Director: Clinical Services, Pretoria Academic Hospital

DR R SOMMERS; MBChB; M.Med (Int); MPhar.Med.
SECRETARIAT of the Faculty of Health Sciences Research Ethics Committee - University of Pretoria

* = Members attended the meeting on 6/09/2006.

141

Appendix III: Ethical approval from the Department of Health, for the genetic screening of BRCA1 mutation is black patients with breast cancer

DEPARTMENT DEPARTEMENT
OF VAN
HEALTH GESONDHEID

Tel:
(012) 354 1560
Fax/Faks:
(012) 354 1831
Ref/Verw:
Ethics Committee
Enquiries/Navrae:
Dr R Sommers
Ward 4 Room 19

Date : 30/06/1999

Number	:	Addendum to KCT 265
Title	:	Extension of Genetic Study: "Screening for specific BRCA1-Mutations in Black Patients Affected With Breast Cancer.
Investigator	:	Prof E J van Rensburg; Dept of Human Genetics; Pretoria Academic Hospital; Pretoria.

This Addendum has been considered by the Ethics Committee,Faculty of Medicine,Univ.of Pretoria and Pretoria Academic Hospitals on 30/06/1999 and found to be acceptable.

Dr J.E.Davel	(female) MBChB:Hospital Superintendent
Prof C.I. Falkson	(female)MBChB;M.Med(Int);MD;Med. Oncologist
Prof G. Falkson	CHAIRPERSON; MBChB;M.Med(Int);MD;OSG: Medical Oncologist
Prof S.V. Grey	(female)BSc(Hons)(Stell);MSc(PU vir CHO) DSc(Pret): Deputy Dean
Mrs C Gerber	(female)BA(FineArts)(U.P)(Unisa);Architectural Draughting (Bostom House College Pta)
Dr V.O.L. Karusseit	MBChB;MFGP(SA);M.Med(Chir); FCS (SA): Surgeon
Dr S.Khan	(female)MB.BCh.(Rand); Med.Adviser (Gauteng Dept.of Health).
Ms B.C.F.Magardie	(female) BCur:Matron/Senior Nursing-Sister
Snr Sr J. Moerane	(female) BCur(Et.Al)Senior Nursing-Sister
Dr P.Z Njongwe	(female)MBChB(Natal);Chief Med.Super of Pretoria Academic Hospital
Prof H.W. Pretorius	MBChB;M.Med (Psych) MD: Psychiatrist
Prof P. Rheeder	MBChB;MMed(Int);LKI(SA);MSc (KLIN.EPI):Specialist Physician
Prof J.R. Snyman	MBChB,M.Pharm.Med: MD:Pharmacologist
Prof De K.Sommers	BChB; HDD; MBChB; MD: Pharmacologist
Dr R Sommers	SECRETARIAT (female)MBChB; M.med (Int); MPhar.Med;
Prof FFW van Oosten	BA; LLB(Potch); LLD (Pret);LLD (Unisa):Head of Department of Public Law and Prof in Criminal Law and Medical Law

PROF G FALKSON ; MBChB; M.Med (Int);MD;
CHAIRPERSON

Pretoria Academic Hospital Private Bag X169 Pretoria 0001 • Pretoria Akademiese Hospitaal Privaatsak X169 Pretoria 0001

Appendix IV: Ethical approval from Ethical approval from the Research Ethics Committee, Faculty of Health Sciences, University of Pretoria for the study investigating prostate cancer risk in indigenous African population.

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

❋ **FWA** 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.

❋ **IRB** 0000 2235 IORG0001762 Approved dd 13/04/2011 and Expires 13/04/2014.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Faculty of Health Sciences Research Ethics Committee
Fakulteit Gesondheidswetenskappe Navorsingsetiekkomitee

DATE: 31/05/2012

AMENDMENT	1. We had to make the following modifications:
	  ○ Updated the list of co-workers
	  ○ Increased the study numbers.
	  ○ Changed the affiliation of the US PI Dr Vanessa Hayes.
	  ○ Included the JCVI ethics approval.
	
	2. Permission requested for expedited approval of a modified study protocol that is now under consideration at NIH for funding. It will link to the APCR approved protocol and is entitled 'Impact of mitochondrial genome variation in extreme prostate cancer disparities'.
	
	The consent form of the APCR study anticipated and therefore included specimens collected for medical diagnoses. The proposed addition will result in no more than minimal risk to human subjects and only adds specimens collected for medical diagnoses. **Application A-G.**

PROTOCOL NO.	**43/2010**
PROTOCOL TITLE	Genetic factors of prostate cancer risk in indigenous African population groups.
INVESTIGATOR	**Principal Investigator:** Prof M S Bornman
SUBINVESTIGATOR	Dr V Hayes
SUPERVISOR	Dr EM Moshokoa
DEPARTMENT	**Dept:** Urology **E-Mail:** mbornman@medic.up.ac.za / marie.odendaal@up.ac.za
STUDY DEGREE	None
SPONSOR	None
MEETING DATE	30/05/2012

The Protocol Amendment (1 and 2 - see above list) were approved on 30/05/2012 by a properly constituted meeting of the Ethics Committee.

Members of the Research Ethics Committee:

Prof M J Bester	(female)BSc (Chemistry and Biochemistry); BSc (Hons)(Biochemistry); MSc(Biochemistry); PhD (Medical Biochemistry)
Prof R Delport	(female)BA et Scien, B Curationis (Hons) (Intensive care Nursing), M Sc (Physiology), PhD (Medicine), M Ed Computer Assisted Education
Prof JA Ker	MBChB; MMed(Int); MD – Vice-Dean (ex officio)
Dr NK Likibi	MBB HM – Representing Gauteng Department of Health) MPH
Dr MP Mathebula	(female)Deputy CEO: Steve Biko Academic Hospital; MBCHB, PDM, HM
Prof A Nienaber	(female) BA(Hons)(Wits); LLB; LLM; LLD(UP); PhD; Dipl.Datametrics(UNISA) – Legal advisor
Mrs MC Nzeku	(female) BSc(NUL); MSc(Biochem)(UCL, UK) – Community representative
Prof L M Ntlhe	MbChB (Natal) FCS (SA)
Snr Sr J Phatoli	(female) BCur(Eet.A); BTec(Oncology Nursing Science) – Nursing representative
Dr R Reynders	MBChB (Prêt), FCPaed (CMSA) MRCPCH (Lon) Cert Med. Onc (CMSA)
Dr T Rossouw	(female) MBChB (cum laude); M.Phil (Applied Ethics) (cum laude), MPH (Biostatistics and Epidemiology (cum laude), D.Phil

Dr L Schoeman	(female) B.Pharm, BA(Hons)(Psych), PhD – Chairperson: Subcommittee for students' research
Mr Y Sikweyiya	MPH; SARETI Fellowship in Research Ethics; SARETI ERCTP;
	BSc(Health Promotion)Postgraduate Dip (Health Promotion) – Community representative
Dr R Sommers	(female) MBChB; MMed(Int); MPharmMed – **Deputy Chairperson**
Prof TJP Swart	BChD, MSc (Odont), MChD (Oral Path), PGCHE – School of Dentistry representative
Prof C W van Staden	MBChB; MMed (Psych); MD; FCPsych; FTCL; UPLM - **Chairperson**

DR R SOMMERS; MBChB; MMed(Int); MPharmMed.
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

Faculty of Health Sciences

16 July 2020

Approval Certificate
Amendment

Ethics Reference No.: 43/2010
Title: DETERMINING GENETIC FACTORS OF PROSTATE CANCER RISK IN FOUR INDIGENOUS AFRICAN POPULATION GROUPS.

Dear Prof MS Bornman

The **Amendment** as supported by documents received between 2020-06-26 and 2020-07-15 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 2020-07-15.

Please note the following about your ethics approval:
- Please remember to use your protocol number (43/2010) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers
MBChB MMed (Int) MPharmMed PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).

Appendix V: Ethical approval from Ethical approval from the Research Ethics Committee, Faculty of Health Sciences, University of Pretoria for the study investigating epigenomic changes due to pesticide exposure in a malaria area

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 22/04/2017.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

16/07/2014

Approval Certificate
Amendment
(to be read in conjunction with the main approval certificate)

Ethics Reference No.: 43/2003

Title: Disruption of the paternal epigenome due to pesticide exposure in a malaria area –Genetic study

Dear Prof C de Jager / Dr Natalie Aneck-Hahn

The **Amendment** as described in the documents received on 9/07/2014 was Provisional approval, to be ratified by the Faculty of Health Sciences Research Ethics Committee on the 30/07/2014. [Semen sample extra: Informed Consent Version 1 genetics dd July 2014]

Please note the following about your ethics amendment:
- Please remember to use your protocol number (**43/2003**) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committe may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics amendment is subject to the following:
- The ethics approval is conditional on the receipt of 6 monthly written Progress Reports, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

_____
Dr R Sommers; MBChB; MMed (Int); MPharMed.
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health).

♦ Tel:012-3541330   ♦ Fax:012-3541367   Fax2Email: 0866515924   ♦ E-Mail: fhsethics@up.ac.za
♦ Web: //www.healthethics-up.co.za   ♦ H W Snyman Bld (South) Level 2-34   ♦ Private Bag x 323, Arcadia, Pta, S.A., 0007

Appendix VI: TaqMan Genotyper Software detailed analysis.
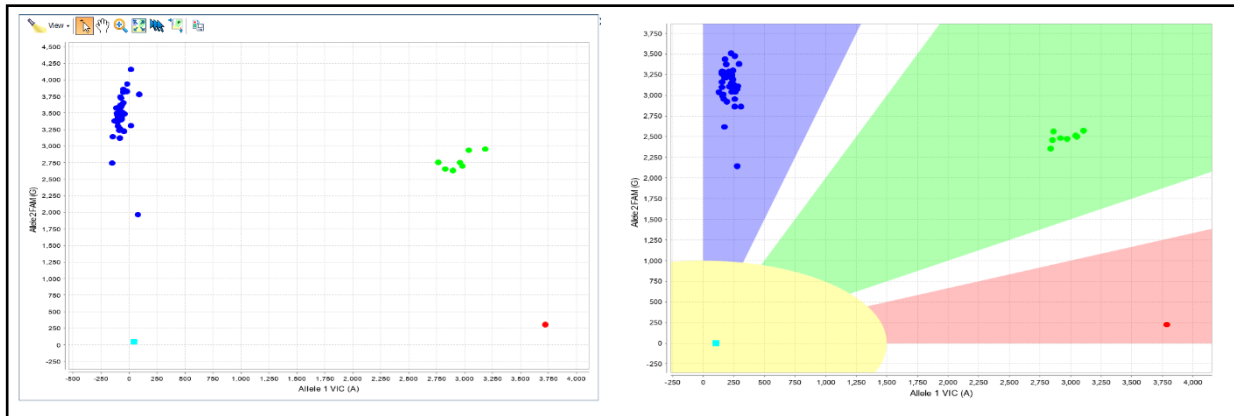
Figure 34. TaqMan® Genotyper Software cluster division, tight clustering due to normalization of DNA to 50 ng/μL. The graph of the left shows the raw data without the clustering. The graph of the right includes the cluster analysis. The red cluster represents samples with the normal sequence for both alleles (VIC/VIC or homozygous for the VIC allele). The green cluster includes individuals having one normal allele and one mutant allele (VIC/FAM or heterozygous). The blue cluster represents homozygous mutants containing both copies of the mutant alleles (FAM/FAM). The yellow region with the one blue squares corresponds to the fluorescence of the non-template control (NTC), showing no amplification. Good data clustering, reduces the potential of genotyping errors.
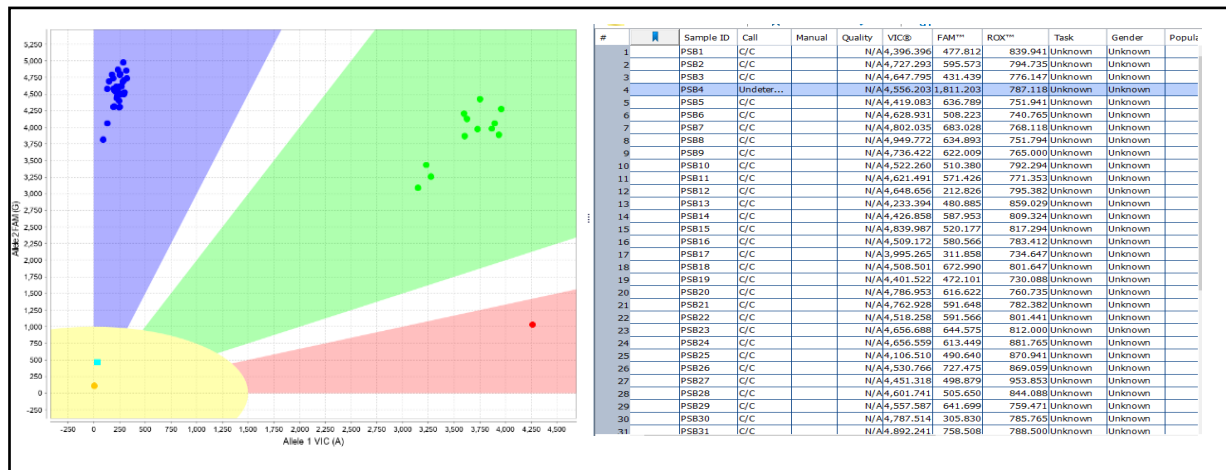
Figure 35. Reflection of samples showing no amplification. The yellow region with the one blue squares corresponds to the fluorescence of the non-template control, the yellow dot corresponds to a sample that showing no amplification. The call on the right-side table is designated "NOAMP". Absence of amplification can be due to poor DNA quality, problems with the automated loaded, altered nucleotide changes and gene deletions. Unfortunately, the exact cause of the lack of amplification cannot be elucidated unless targeted sequencing is performed.
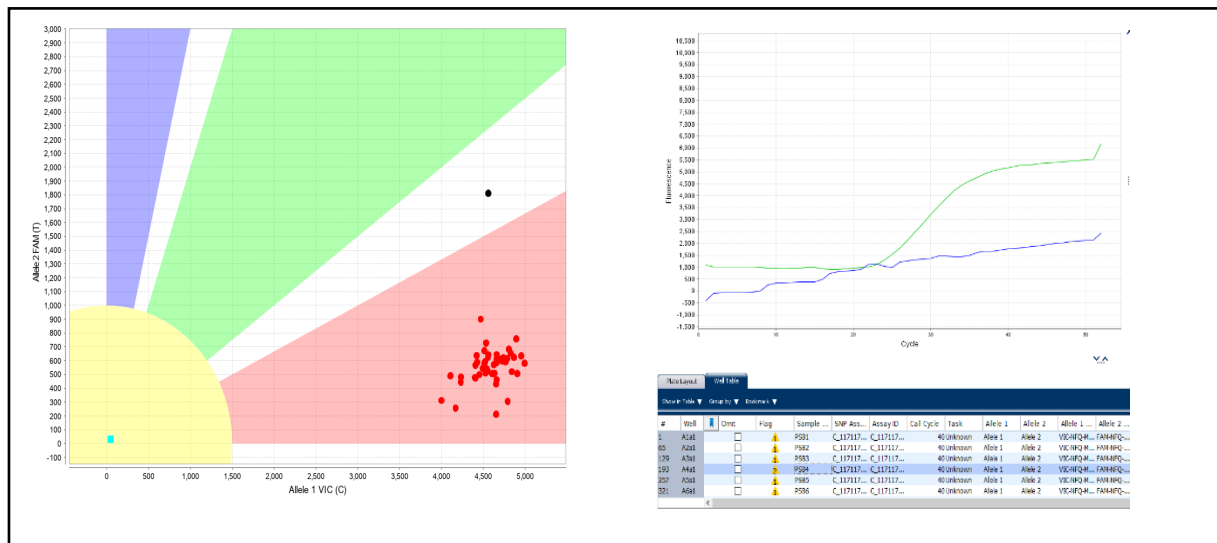
Figure 36. Automatic computer determination of genotype unable to determine genotype (observed as a black dot). In these events the QuantStudio® 12K Flex Software, is used to confirm which MGB probe amplified (VIC or FAM or both). In the above example, the genotype will be designated as homozygous for the VIC or C allele.
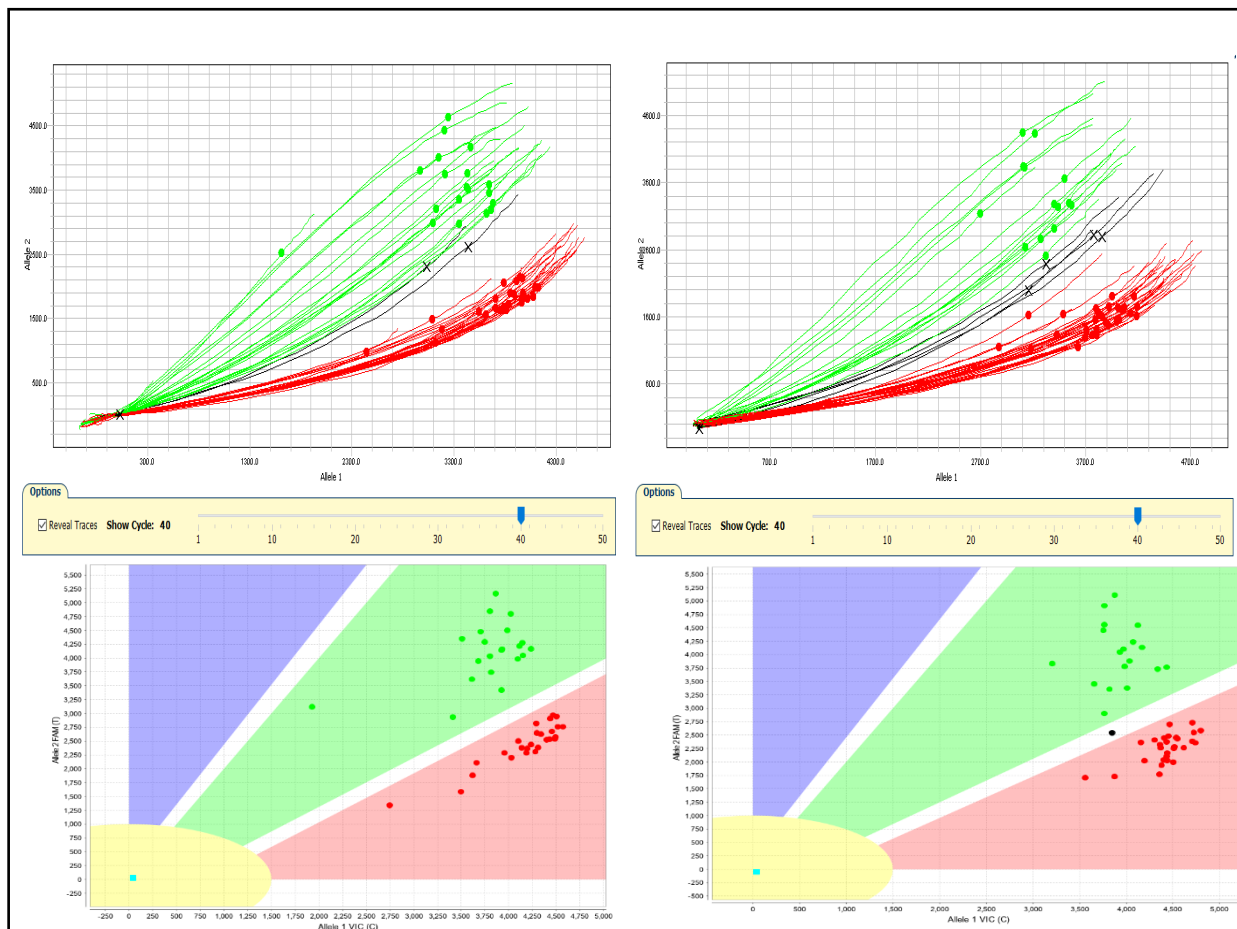
Figure 37. Comparison between QuantStudio® 12K Flex Software (top) and TaqMan® Genotyper Software (bottom). The classification tool on the TaqMan® Genotyper Software improves the clustering. The black dot seen in the TaqMan® Genotyper Software analysis bottom right, needs to be determined using the discretion of the analyst.
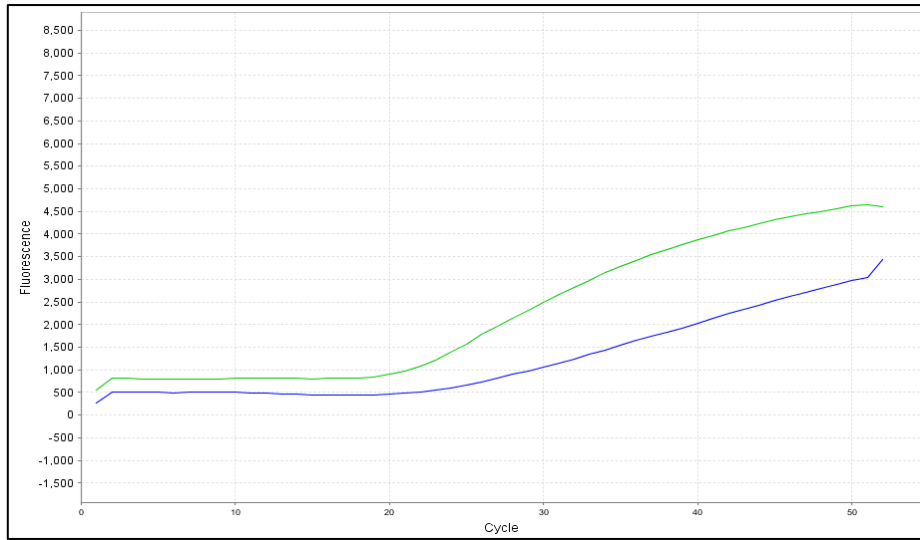
Figure 38. Multicomponent plot QuantStudio® 12K Flex, acts as a visual aid reflecting the amplification of both the VIC (green) and FAM (blue) probe, indicative of the sample being heterozygous for the alleles.

Appendix VII: Pairwise sequence analysis

primates | 62 leaves

Homo sapiens FOSMID clone ABC9-43905700J11 from chromosome 22, complete sequence
Homo sapiens cytochrome P450 2D7 (CYP2D7P) pseudogene, partial sequence; and cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*36 allele, partial cds
Homo sapiens cytochrome P450 2D7 (CYP2D7P) pseudogene, partial sequence; and cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*10 allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*36 allele and cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*10 allele, complete cds
Homo sapiens cytochrome P4502D6 (CYP2D6) gene, CYP2D6*1 allele, complete cds
Homo sapiens premature mRNA for debrisoquine 4-hydroxylase mutant allele variant
Human DNA sequence from clone RP4-669P10 on chromosome 22q13.31-13.33, complete sequence
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*58 allele, complete cds
Homo sapiens nonfunctional cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*56 allele, complete sequence
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*41 allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*17V allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*17 allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*2D allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*45B allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*29 allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*10B allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*9 allele, complete cds
Homo sapiens nonfunctional cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*4D allele, complete sequence
Homo sapiens nonfunctional cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*4A allele, complete sequence
Homo sapiens nonfunctional cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*3 allele, complete sequence
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*2L allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*1V_AA allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*1_AA allele, complete cds
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*1V allele, complete cds
Human debrisoquine 4-hydroxylase mutant allele (CYP2D6-MA1) gene, complete cds
Human cytochrome P450 IID6 (CYP2D6) gene, complete cds
Human mRNA for cytochrome P-450IID (clone pMP34)
Human CYP2D7BP pseudogene for cytochrome P450 2D6
lcl|Query_48073

Figure 39. Pairwise alignment distance tree is used to compare the relatedness of sequences. Samples that are very closely related have no distance between the nodes and align in a straight line. The above pairwise alignment tree for *CYP2D6*10*, depicts 100% homology to other several gene regions, including other gene regions of *CYP2D6*.

Figure 40. Pairwise alignment distance tree is used to compare the relatedness of sequences. Samples that are very closely related have no distance between the nodes and align in a straight line. The above pairwise alignment tree for *CYP2D6*4*, depicts 100% homology to other several gene regions, including other gene regions of *CYP2D6*.
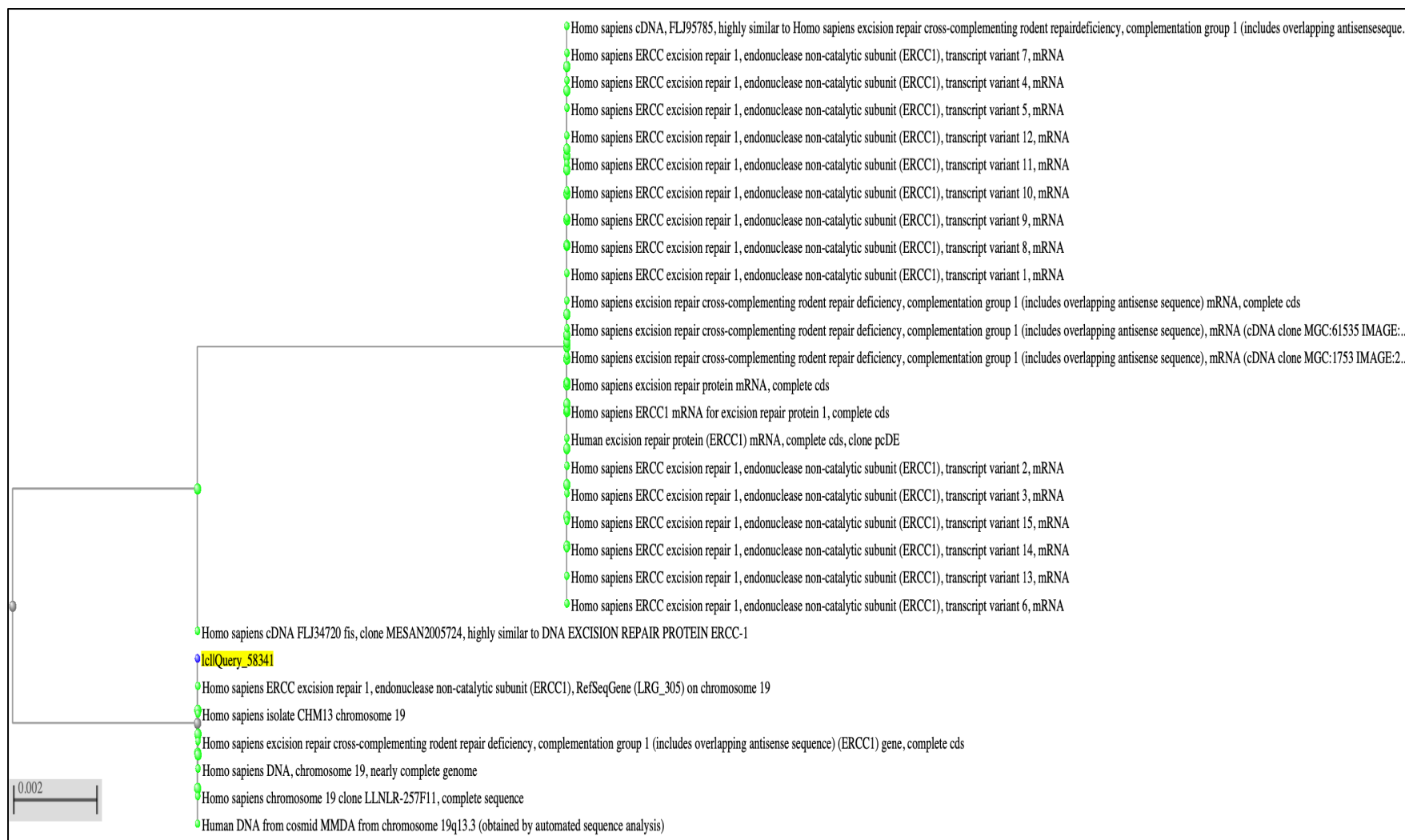
Figure 41. Pairwise alignment distance tree *CYP2D6*3*, depicting the 100% homology to other *CYP2D6* gene regions.

Homo sapiens BAC clone RP11-65I14 from chromosome 22, complete sequence
Homo sapiens FOSMID clone ABC10-44146800G7 from chromosome 22, complete sequence
Homo sapiens FOSMID clone WI2-2025H20 from chromosome 22, complete sequence
Homo sapiens FOSMID clone WI2-3087P5 from chromosome 22, complete sequence
Homo sapiens FOSMID clone ABC9-44010900K17 from chromosome 22, complete sequence
Homo sapiens BAC clone CH17-256B8 from chromosome 22, complete sequence
Homo sapiens FOSMID clone ABC14-50078100O11 from chromosome 22, complete sequence
Homo sapiens FOSMID clone ABC12-49046000E12 from chromosome 22, complete sequence
Homo sapiens FOSMID clone WI2-3087L19 from chromosome 22, complete sequence
Homo sapiens isolate CHM13 chromosome 22
Homo sapiens DNA, chromosome 22, nearly complete genome
lcl|Query_21209
primates | 73 leaves
Homo sapiens haplotype Baka09b cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Baka09a cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Baka07b cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Baka05a cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Baka04b cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Bkk13a cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Bkk12b cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Bkk12a cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Bkk06b cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens haplotype Bkk02b cytochrome P450 2D6 (CYP2D6) gene, complete cds
Homo sapiens nonfunctional cytochrome P450 family 2 subfamily D polypeptide 6 (CYP2D6) gene, CYP2D6*13A2 allele, complete cds
Homo sapiens cytochrome P450, family 2, subfamily D, polypeptide 6 (CYP2D6) gene, complete cds
Homo sapiens CYP2D7/CYP2D6 fusion protein (CYP2D6*13A2) gene, partial sequence
Homo sapiens CYP2D6/CYP2D7 fusion protein (CYP2D6*63) gene, complete sequence
Homo sapiens CYP2D6/CYP2D7 fusion protein (CYP2D6*61) gene, complete sequence
Homo sapiens cytochrome P450 2D6 (CYP2D6) gene, CYP2D6*2A variant allele, complete cds

0.002

Figure 42. Pairwise alignment distance tree *CYP2D6*17*, depicting the 100% homology to other *CYP2D6* gene regions.

154

Figure 43. Pairwise alignment distance tree shows the sequences of similarity by the distance between the nodes. The above pairwise alignment tree is that of *ERCC1*, depicting the 100% homology to the CHM13 gene on chromosome 19.

155
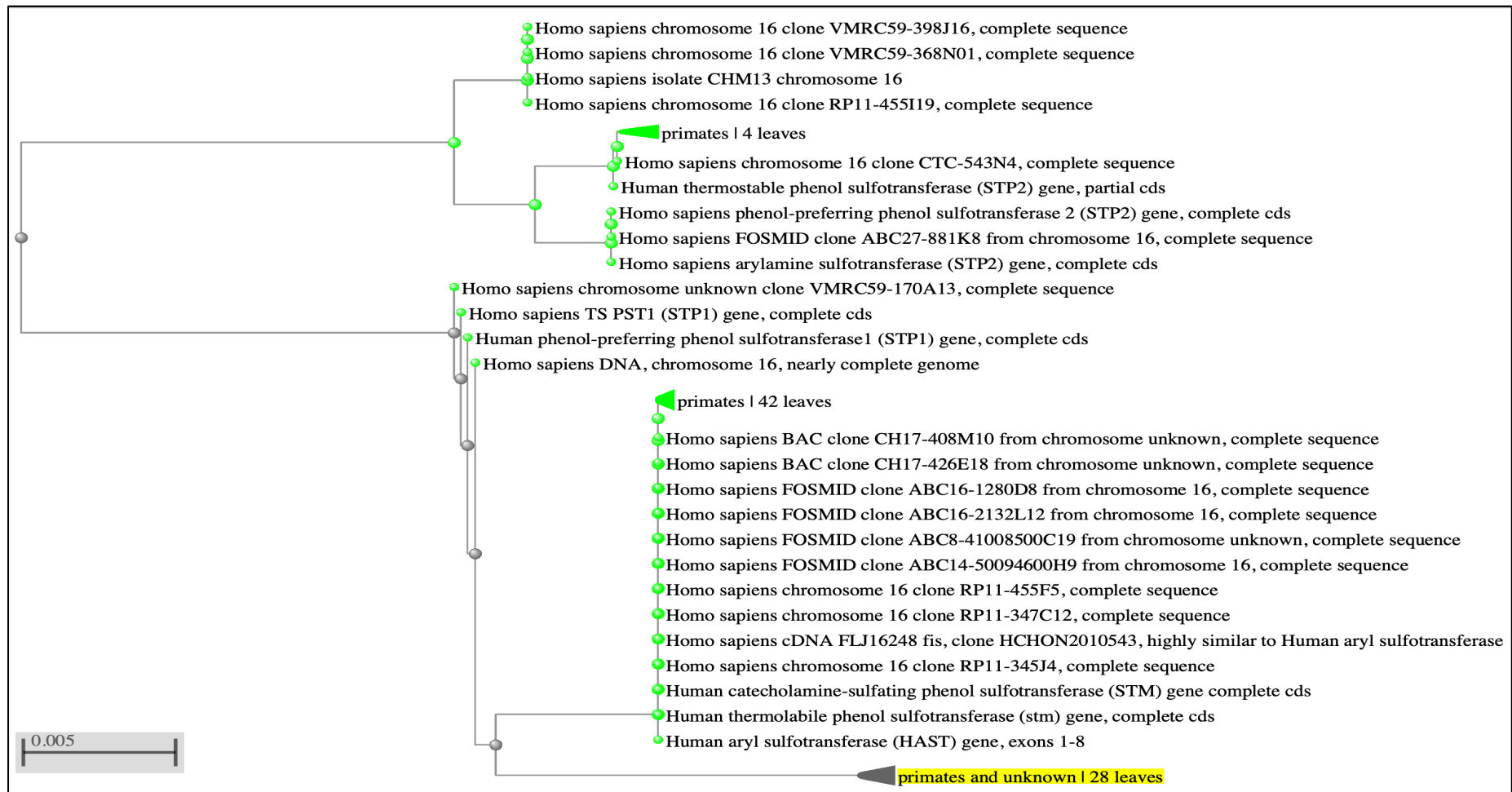
Figure 44. Pairwise alignment distance tree shows the sequences of similarity by the distance between the nodes. The above pairwise alignment tree is that of *SULT1A1*2*, depicting the 100% homology to the *STP1* gene.

Appendix VIII: Hardy-Weinberg Equilibrium test

1. HWE Test with cancer cohorts removed: *SULT1A1*2* (rs1042028)

A simple calculator to determine whether observed genotype
frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	*Observed #	Expected #
Homozygote reference:	240	258,5
Heterozygote:	179	142,0
Homozygote variant:	1	19,5
Var allele freq:	0,22	420

Chi-squared value = 28,5174887
Chi-squared test P value = 0,000000
(if < 0.05 - not consistent with HWE)

2. HWE Test with cancer cohorts removed: *ERCC1* (rs11615)

A simple calculator to determine whether observed genotype
frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	*Observed #	Expected #
Homozygote reference:	385	381,8
Heterozygote:	29	35,4
Homozygote variant:	4	0,8
Var allele freq:	0,04	418

Chi-squared value = 13,53129168
Chi-squared test P value = 0,000235
(if < 0.05 - not consistent with HWE)

3. HWE Test with cancer cohorts removed: *CYP2D6*3* (rs35742686)

A simple calculator to determine whether observed genotype
frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	*Observed #	Expected #
Homozygote reference:	412	411,0
Heterozygote:	0	2,0
Homozygote variant:	1	0,0
Var allele freq:	0,00	413

Chi-squared value = 413

Chi-squared test P value = 0,000000

(if < 0.05 - not consistent with HWE)

4. HWE Test with cancer cohorts removed: *CYP2D6*4* (rs3892097)

A simple calculator to determine whether observed genotype
frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	*Observed #	Expected #
Homozygote reference:	395	392,3
Heterozygote:	17	22,4
Homozygote variant:	3	0,3
Var allele freq:	0,03	415

Chi-squared value = 23,86495082

Chi-squared test P value = 0,000001

(if < 0.05 - not consistent with HWE)

5. HWE Test with cancer cohorts removed: *CYP2D6*10* (rs1065852)

A simple calculator to determine whether observed genotype
frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	*Observed #	Expected #
Homozygote reference:	349	337,8
Heterozygote:	49	71,4
Homozygote variant:	15	3,8
Var allele freq:	0,10	413

Chi-squared value = 40,75935232

Chi-squared test P value = 0,000000

(if < 0.05 - not consistent with HWE)

6. HWE Test with cancer cohorts removed: *CYP2D6*17* (rs28371706)

A simple calculator to determine whether observed genotype
frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	*Observed #	Expected #
Homozygote reference:	261	246,5
Heterozygote:	113	142,1
Homozygote variant:	35	20,5
Var allele freq:	0,22	409

Chi-squared value = 17,11467207

Chi-squared test P value = 0,000035

(if < 0.05 - not consistent with HWE)