UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

IDENTIFICATION OF EXPRESSION QUANTITATIVE TRAIT LOCI (eQTLs) AND THEIR FUNCTIONAL IMPACT ON ADME GENES

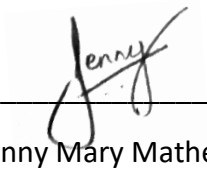
Jenny Mary Mathew

A Dissertation submitted to the Faculty of Science, University of the Witwatersrand, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, 2021

DECLARATION

I, Jenny Mary Mathew declare that this Thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



_____

Jenny Mary Mathew

on 1st November 2021 at Parktown, Johannesburg

DEDICATION

This thesis is dedicated to my husband Siju Abraham Mammen. Thank you for your love and for having been with me throughout this journey. It has been a bumpy ride and I am aware that I did not make it any easy on you, but without your encouragement, hugs and coffee-pick-me-ups, this would not have been possible.

To my daughters Ava and Zaya, your cuddles, kisses and kicks will forever be treasured. I hope I inspire you both to aim high and shoot for the stars.

To my parents, Letha and Thampy, thank you for having raised me the way you did, for instilling in me that giving up is not an option. To my brothers, Jerrish and Jebby, I hope I have made both of you proud. Thank you for your love.

To all my family and friends, thank you for your love, support and encouragement. Thank you for holding me accountable and for always believing in me.

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to the following people and institutional bodies who have contributed towards the completion of this PhD:

- My supervisors Prof. Michèle Ramsay and Prof. Demetra Mavri Damelin. Thank you for your guidance and motivation without which this journey would have been impossible.
- Prof. Scott Hazelhurst, for forwarding my CV to Prof Ramsay without which a PhD would have just remained a dream.
- Dr. Phelelani T. Mpangase, Dr Dhriti Sengupta and Dr Stanford Kwenda, friends, mentors, for your guidance and for graciously helping me out throughout the course of this PhD. Thank you for being tolerant to all my doubts and questions.
- My friends at SBIMB, especially Mahtaab Hayat, Shaun Aron, Evans Mathebula, Jorge Da Rocha, Carl Chen, Romuald Boua, Hleli Mthimkhulu and Ananyo Chaudhary. Thank you for checking in on me throughout this journey and for your kind words of encouragement.
- The University of the Witwatersrand and the National Research Foundation (NRF) for providing the funding which made it possible to complete this work.

ABSTRACT

Polymorphisms in absorption, distribution, metabolism and excretion (ADME) genes result in inter-individual variation in drug response, drug efficacy and adverse drug effects caused by drug toxicity. Expression quantitative trait loci (eQTLs) are genetic variants that are involved in gene expression regulation that may result in phenotype variation. In the context of this study, the phenotype being studied is variation in drug response. ADME gene-associated eQTLs can help predict the clinical outcome of treatment as they affect the expression of ADME genes that modulate drug metabolism and transport. eQTLs play an important role in pharmacogenomics as it is easier and less costly to determine an individual's genotype as a proxy for gene expression than to isolate RNA or protein to guide drug choice and dose best suited to an individual's genetic makeup. A robust bioinformatic pipeline was developed that identifies tissue-specific eQTLs (https://github.com/phelelani/nf-eqtl). We focussed on identifying eQTLs that affect ADME gene expression in the liver. The pipeline identified cis-eQTLs for six core ADME genes (*CYP2C19, CYP3A5, DPYD, SULT1A1, UGT1A1* and *UGT2B17*). Though the pipeline identified cis-eQTLs for 63 extended ADME genes, only *UGT1A6, UGT2B15, ABCC3, ABCC4, XDH* and *CES1* were involved in the metabolism of drugs commonly prescribed in South Africa (SA) for human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS), tuberculosis, diabetes, hypertension, pain, and cardiovascular diseases. The cis-eQTLs identified in this study were found to be present at very different frequencies in African populations compared to other populations. This highlights the importance of doing more eQTL studies with data from individuals of African origin to understand differences pertaining to pharmacogenetics. eQTLs identified in this study are hypothesized to predict the level of protein produced and therefore potential clinical outcomes of drug treatment. If validated, treatment response predictions/algorithms can be enhanced by adding these eQTLs in combination with existing knowledge of genetic variations and other clinical factors. Developing pharmacogenetic tests that include eQTLs in addition to existing well-known genetic variants can be a novel strategy to optimise treatment and identify individuals who may be at risk for adverse drug events as well as those who are more likely to achieve a therapeutic response or require alternative treatment.

RESEARCH OUTPUTS

Mathew JM, Mpangase P, Sengupta D, Kwenda S, Mavri-Damelin D, Ramsay M (2021) 'UGT1A1 regulatory variant with potential effect on efficacy of HIV and cancer drugs commonly prescribed in South Africa ', Pharmacogenomics, *in press*

The Journal Paper is attached in Appendix C.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

5'-DFCR	5'-deoxy-S-fluorocytidine
5'-DFUR	5'-deoxy-S-fluorouridine
5-FU	Fluorouracil
6-MP	6-Mercaptourine
ADME	Absorption, Distribution, Metabolism and Excretion
ADR	Adverse drug reaction
AFR	African
AGVP	African Genome Variation Project
AIDS	Acquired immunodeficiency syndrome
ASW	Americans of African Ancestry in Southwest USA
AUC	Area under the concentration-time curve
AZA	Azathioprine
BCRP	Breast cancer resistance protein
BIG	Brain Imaging Genetic
CCR5	Chemokine C-C motif receptor 5
CEU	Utah Residents with Northern and Western European Ancestry from the CEPH collection
CML	Chronic myeloid leukaemia
CNP	Copy number polymorphism
CNV	Copy Number variant
COMT	Catechol O-methyl transferase
CRC	Colorectal cancer
CYP	Cytochrome P450
Cyt D	Cytidine deaminase
DME	Drug metabolising enzyme
DPD	Dihydropyrimidine dehydrogenase
DSL	Domain specific language
EM	Extensive metaboliser
EMA	European Medicines Agency
eQTL	Expression quantitative trait loci
eSNPs	Expression single nucleotide polymorphisms
EUR	European
FDR	False Discovery Rate
G6PD	Glucose-6-phosphate dehydrogenase
GAD	Generalized anxiety disorder
GEO	Gene Expression Omnibus
GERD	Gastroesophageal reflux disease
GIST	Gastrointestinal stromal tumor
GST	Glutathione-S-transferase
GWAS	Genome wide association studies
H/K ATPase	Hydrogen/potassium adenosine triphosphatase
HCl	Hydrochloric acid
HCSC	Health Canada/Santé Canada

HIV	Human immunodeficiency virus
HWE	Hardy Weinberg equilibrium
IM	Intermediate metaboliser
IV	Intravenous
Kb	Kilobase
KGP	1000 Genomes Project
LD	Linkage disequilibrium
MaCH	Markov Chain Haplotyping
MAF	Minor allele frequency
Mb	Mega base
MDD	Major depressive disorder
NADPH	Nicotinamide adenine dinucleotide phosphate
NAT	N-acetyltransferase
NCBI	The National Center for Biotechnology Information
NIH	National Institutes of Health
NME	New molecular entity
NPAEs	Neuropsychiatric adverse events
OCD	Obsessive compulsive disorder
OS	Overall survival
PCA	Principal component analysis
PFS	Progression free survival
PharmGKB	The Pharmacogenomics Knowledge Base
PM	Poor metaboliser
PMDA	Pharmaceuticals and Medical Devices Agency
PPI	Proton pump inhibitor
QC	Quality control
QTL	Quantitative trait loci
RIN	RNA Integrity Number
SA	South Africa
SEB	South eastern Bantu speakers
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOP	Standard operating procedure
SSRI	Selective serotonin reuptake inhibitor
SULT	Sulfotransferase
SULT1A1	Sulphotransferase 1A1
Swissmedic	Swiss Agency of Therapeutic Products
TAP	Transporter associated with antigen processing
TDM	Therapeutic drug monitoring
TP	Thymidine phosphorylase
TPM	Transcripts per million
TPMT	Thiopurine methyltransferase
TPMT	Thiopurine S-methyl transferase
TSS	Transcription start site
UGT	Uridine 5'-diphospho-glucuronosyltransferases
UGT	Glucuronosyltransferase

UGT1A1	Uridine diphosphate glucuronosyl transferase1A1
UK	United Kingdom
UM	Ultrarapid metaboliser
US FDA	United States Food and Drug Administration
VEP	Ensembl Variant Effect Predictor
VIP	Very Important Pharmacogene
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization
YRI	Yoruba in Ibadan, Nigeria

CHAPTER 1 INTRODUCTION/BACKGROUND

In recent years it has become possible for genetic studies to move from variant genotyping to the analysis of genome function because of high-throughput technologies. The most commonly explored link in this context is the contribution of genetic variation to gene expression, commonly referred to as expression quantitative trait loci (eQTL) (Börnigen *et al.*, 2018; Cao, 2020). eQTLs are defined as loci in the genome that correlate with variation in gene expression (Smith, 2020). An eQTL can potentially explain a fraction of the effect that genetic variance has on gene expression (Nica *et al.*, 2013). Since these chromosomal loci modulate gene expression, the effect that genetic variation has on gene expression can be assessed by eQTLs. This helps to improve our understanding of how genetic variation can result in phenotypic variation by linking genetic variation to gene expression modulation (Jung *et al.*, 2020). eQTLs are primarily investigated to gain insight into the characterization of genetic variation. They help in the better understanding and interpretation of gene regulation and genome-wide association studies (Nica *et al.*, 2013).

In the context of pharmacogenomics, eQTLs have the intrinsic potential to assist in the prediction of the clinical outcome of treatment as they are associated with changes in the expression of genes that modulate drug metabolism and transport, such as the absorption, distribution, metabolism and excretion (ADME) genes. eQTL analysis is a powerful genetic tool to assess whether single nucleotide polymorphisms (SNPs) affect phenotype by gene expression modulation. This can be used to predict and assess how genetic variants, like SNPs, affect drug efficacy (Hovelson *et al.*, 2017). eQTLs therefore play an important role in pharmacogenomics as it is both economically and practically more feasible to determine an individual's genotype as a proxy for gene expression, to ensure that a patient is prescribed a drug at a dose best suited to their genetic makeup. Since eQTLs are SNPs that modulate gene expression in specific tissues, they can help uncover genetic factors and biochemical pathways that cause diseases (Shabalin, 2012).

1.1 WHAT ARE ADME GENES?

ADME genes are involved in the absorption, distribution, metabolism, and excretion of drugs. They encode Phase I and II drug metabolising enzymes, transporters, and modifiers.

ADME genes play a critical role in the efficacy of drugs through determining how much drug has been absorbed, distributed, metabolised and excreted, and hence control drug levels in the body (Hovelson *et al.*, 2017). ADME genes therefore play a vital role in how the body responds to a drug (pharmacokinetics) and how the drug affects the body (pharmacodynamics). Pharmacokinetics is defined as, "the study of the rate and extent of drug absorption, distribution, metabolism, and excretion" (Belle *et al.*, 2008). Thus, pharmacokinetics determines the fate of a drug in the body. Pharmacodynamics is defined as the study of what the drug does to the body (Belle *et al.*, 2008). ADME genes thus influence the pharmacological activity of the drug.

There is considerable genetic variation in ADME genes (Hovelson *et al.*, 2017). These variations affect drug pharmacokinetics and pharmacodynamics, either by altering the structure and function of the encoded protein, by modifying the expression level of the protein it encodes, or by altering the activity of the enzyme, all of which contribute significantly to individual patients' drug sensitivity, resistance and toxicity (Arbitrio *et al.*, 2016)**.** Data on genetic variation in ADME genes can help reduce adverse drug reactions (ADRs) by helping researchers and clinicians understand underlying causes for sensitivity/resistance to drugs (Iskakova *et al.*, 2016).

The liver is the primary organ of xenobiotic metabolism and clearance. A xenobiotic is a chemical substance foreign to the human body. It includes plant constituents like pollen, drugs, pesticides, cosmetics, flavourings, fragrances, food additives, industrial chemicals and environmental pollutants (Patterson *et al.*, 2010). Though many organs like the brain and oesophagus express some ADME genes like *CYP2B6*, *CYP2D6*, *CYP2A6*, *CPY2A7*, *CYP2E1* (Godoy *et al.*, 2002; Ferguson *et al.*, 2011), the liver is the major detoxifying organ. Therefore, the liver expresses a variety of drug metabolising enzymes and drug transporters. (Hu *et al.*, 2019).

Drug metabolising enzymes and drug transporters (DMET) help in the distribution of drugs in the body. These can be classified into three broad categories. The first category is phase I enzymes (Ekhart, Rodenhuis, *et al.*, 2009). Phase I enzymes help in the transformation of a drug to more polar metabolites by means of oxidative, reductive, or hydrolytic reactions that usually create or reveal a functional group (e.g. hydroxyl, carboxyl, amino) on a xenobiotic (Jancova *et al.*, 2010), as carried out by enzymes including hydrolases, reductases and oxidases. An example of a Phase I enzyme is the cytochrome P450 family of enzymes (CYPs). Cytochromes P450 constitutes a superfamily of haem enzymes responsible for the metabolism of xenobiotics. Most drugs are metabolised by CYPs either as a route to detoxification or as an activation pathway for an inactive prodrug. A prodrug is a drug that is pharmacologically inactive (Vemula *et al.*, 2011; Preissner *et al.*, 2015). To exhibit therapeutic efficacy, a prodrug must be activated via chemical reactions in the body like hydrolysis or phosphorylation into active metabolites (Vemula *et al.*, 2011; Preissner *et al.*, 2015).

The second category of drug metabolising enzymes is phase II enzymes (Ekhart, Rodenhuis, *et al.*, 2009). Phase II biotransformation reactions are also called conjugation reactions. Phase II enzymes conjugate phase I products, thereby serving as a detoxifying step. They either conjugate the parent compound or reactive intermediates with water-soluble moieties such as glucuronic acid, glutathione or sulphate. The increased water solubility of the drug facilitates excretion into bile/urine. Thus, phase II enzymes play a major role in the metabolic inactivation of pharmacologically active compounds. As such, these enzymes contribute to the intracellular metabolism of many intermediate drug substrates and include UDP-glucuronosyltransferases (UGTs), glutathione-S-transferases (GSTs), sulfotransferases (SULTs), N-acetyltransferases (NATs), and methyltransferases like thiopurine S-methyl transferase (TPMT) and catechol O-methyl transferase (COMT) (Jancova *et al.*, 2010).

The third category is the drug transporters, of which there are two types: influx drug transporters and efflux drug transporters (Ekhart, Rodenhuis, *et al.*, 2009). Drug transporters

are membrane-bound proteins that control the uptake and excretion of drugs and xenobiotics (Playing *et al.*, 2014) into and out of cells. The distribution and the bioavailability of drugs are influenced by drug transporters. Genes that encode drug transporters include the ABC and SLC superfamilies; *MDR1* (*ABCB1*), which encodes P-glycoprotein, and *ABCG2*, which encodes breast cancer resistance protein (BCRP).

The PharmaADME Consortium has currently classified 299 genes as ADME genes. The ADME genes are further categorised as core ADME genes and extended ADME genes (http://www.pharmaadme.org), although there is not universal consensus about this classification. The core ADME genes represent the most important genes that are directly involved in the metabolism and clearance of drugs. The extended ADME gene set refers to other genes that are related to drug metabolism and/or clearance. Currently there are 32 core ADME genes and 267 extended ADME genes. The combined activities of the ADME genes influences the efficacy of drug treatment.

1.2 TYPES OF GENETIC POLYMORPHISMS AND THEIR RELEVANCE IN ADME GENES

The human genome has different genetic variations ranging from large structural chromosome anomalies that are microscopically visible, to changes in single nucleotide bases. Some other genetic variations found in the human genome are substitutions, deletions, duplications, insertions, and complex multiple-site genetic variants, collectively called copy number variants (CNVs) or copy number polymorphisms (CNPs).

A single nucleotide polymorphism (SNP) is the presence of a variant at a single position of the DNA sequence. Allele frequency refers to the prevalence of a genetic variant in a population. The term polymorphism refers to variant alleles that reach a frequency of 1% or above among individuals of a population (Johnston *et al.*, 2019). To be strictly correct, if they are not polymorphic, they should be referred to as single nucleotide variants (SNVs), although they are used interchangeably by the scientific community. This variation in sequence occurs across the entire genome and accounts for well-defined individual phenotypic characteristics

or may have no phenotypic effect. There are roughly four to five million SNPs in an individual's genome making it the most common type of genetic variation. Most of these SNPs are harmless and have no effect on health or development. However, certain SNPs are significant in the study of human health, diseases and susceptibility to environmental factors. The non-random association of alleles from different genetic loci, resulting in alleles being inherited together as a haplotype block, is referred to as linkage disequilibrium (LD) (Slatkin, 2008; Single *et al.*, 2016).

A copy number variant (CNV) is a genetic polymorphism that is defined as a DNA segment of one kilobase (kb) or larger that is present at a variable copy number in comparison with a reference genome (Cooper *et al.*, 2008)**.** CNVs can range from kilobases (kb) to megabases (Mb) in size (Feuk *et al.*, 2006). A CNV can be simple in structure, like a tandem duplication, or can involve complex additions or losses of repetitive sequences at multiple sites in the genome. Some CNVs have no apparent influence on phenotype, while many CNVs have definitive links as predictive biomarkers for clinical outcomes and disease associations(Zhang *et al.*, 2009; Consortium *et al.*, 2010). CNVs have multiple functionalities. They can be functionally neutral, cause disease, confer risk to complex disease traits, or modulate expression of genes causing phenotypic variation.

Polymorphisms in ADME genes have been associated with inter-individual variability in the activity of the ADME genes that can result in variation in drug response, drug efficacy, drug toxicity and disease predisposition (Zhou *et al.*, 2009). SNPs, CNVs and other genetic variants/polymorphisms in ADME genes play clinically important roles in determining enzyme activity and have been shown to alter the metabolism of many drugs. Genetic variations that affect ADME gene expression and/or activity are likely to impact the pharmacokinetics and pharmacodynamics of drug therapy (Huang *et al.*, 2008). There exists considerable genetic variation in ADME genes that can affect the pharmacokinetics of drugs both within a population and between populations. Polymorphisms in drug metabolising enzymes and transporters are of clinical significance because they affect drug safety and efficacy (Ch Yiannakopoulou, 2013).

Variability in drug response between individuals is a major problem in clinical practice and drug therapy, drug development and drug regulation (Zhou *et al.*, 2009). The medical and economic consequences linked to the lack of therapeutic effect of drugs and high occurrence of adverse drug reactions (ADRs) is dire. This inter-individual variation in drug response and distribution cannot only be explained by factors like age of the patient, co-morbidity, life style, co-medication and renal/liver function but also by genetic variations in drug-metabolising enzymes (DMEs), drug transporters, and drug targets (Ma *et al.*, 2011). These genetic variations are important determinants of drug efficacy and toxicity in patients. To achieve maximum therapeutic efficacy with minimum ADRs, clinicians have to be enabled through evidence-based knowledge to prescribe the right medication in the right dose for an individual, an approach often referred to as precision medicine (Sukasem *et al.*, 2016).

The type, frequency of genetic variation in ADME genes and the effect the variation has can differ in magnitude or can be completely absent depending on the population being studied (Bonifaz-Peña *et al.*, 2014). PharmGKB, a database, documents several such pharmacogenetic biomarkers with evidence that their frequencies differ in different populations (https://www.pharmgkb.org/).

Drug pharmacokinetics is controlled by several intrinsic and extrinsic factors. These factors include variation in ADME genes, that amongst other things, help to arrive at a therapeutically effective and safe drug dosage guidelines (Wilson *et al.*, 2001; S. F. Zhou *et al.*, 2009; Stingl *et al.*, 2014). The pharmacokinetic profile of a drug is largely determined by variations in ADME genes (Hovelson *et al.*, 2017). For example, the United States Food and Drug Administration (US FDA) has updated the product labels of more than 20 psychotropic drugs with information relating to ADME genetic variants that have experimentally been shown to contribute to interindividual pharmacokinetic variability (Stingl *et al.*, 2014).

Populations differ in their genetic architectures. This means that different populations have different variants and distributions of variants that may affect how they react to the same

dose of the drug in a different way. Drug response is highly variable both between different populations and within a population (Li *et al.*, 2011; Hovelson *et al.*, 2017). This means that ADME genes show genetic differentiation between and across populations. For example, at the same dose, a two-fold higher systemic concentration of rosuvastatin, a drug administered for lowering bad cholesterol levels, is observed in individuals of East Asian ancestry than European ancestry. Hence, the recommended dosage of rosuvastatin in East Asians is lower, at 5 mg, which is half of what is generally prescribed for other populations (10 mg) (Lee *et al.*, 2013). Publicly available data from the Human Genome Diversity Project and the Phase III HapMap project were used by Li *et al.* (Li *et al.*, 2011) to explore ADME gene selection patterns across global populations. They reported that greater than expected diversity in ADME genes was observed between populations. Therefore, identifying and most importantly quantifying ADME gene genetic variations that result in inter-population differences in drug efficacy, safety and pharmacokinetics is of prime importance (Hovelson *et al.*, 2017). Understanding genetic differences in ADME genes across global populations will help establish strategies for preclinical and clinical development so that pharmacokinetic data from multicentric clinical studies can be understood and interpreted better.

ADME gene variants studied in most studies suffer from ascertainment bias because these variants were discovered and observed in populations of European ancestry (Hovelson *et al.*, 2017). ADME gene variants show significant variation both within and between different continental populations (Hovelson *et al.*, 2017; Rodrigues *et al.*, 2019). African ancestry populations show the most population differentiation in ADME gene variants and East Asian ancestry populations show the least (Ramos *et al.*, 2014).

The list of clinically important biomarkers associated with drug response and toxicity is growing rapidly (Ramos *et al.*, 2014; Lauschke *et al.*, 2017). US FDA maintains a table of pharmacogenomic biomarkers for drug labels used in the United States of America (USA) (http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm). The UK Medicines and HealthCare Products Agency and the European Medicines Agency have protocols in place to consider pharmacogenetic biomarkers for targeted therapy

and drug safety warnings (Ehmann *et al.*, 2015). Physicians will benefit from these clinically validated pharmacogenetic biomarkers, as they will assist them in the optimisation of drug dose, choice of drug and treatment duration, thereby reducing the occurrence of ADRs (Wang *et al.*, 2011). Pharmacogenomics is a core constituent of personalized medicine. Lack of sufficient clinically actionable data is the biggest limitation in the field. It is currently estimated that more than 90% of drugs prescribed lack valid and predictive genetic biomarkers needed to attain drug efficacy and avoid severe adverse drug reactions (Schwab *et al.*, 2012a).

Ethnically diverse population groups from all over the world are poorly represented (Rodrigues *et al.*, 2019). This leads to poor understanding of human pharmacogenomic variation distribution. The need of the hour is to have a comprehensive understanding of ADME gene variation across various populations of the world with diverse ancestral backgrounds. This understanding will help facilitate the translation of pharmacogenetic data into clinical practice and public health policies across the world leading to better healthcare (Swen *et al.*, 2007; Ventola, 2011; Ramos *et al.*, 2014).

1.3 GENETIC ARCHITECTURE OF POPULATIONS

Differences in genetic diversity between and within populations (presence or absence of variants and their presence at different frequencies) determine population genetic structure. Factors that contribute to variations in allele frequencies between populations include geographic isolation or separation, population migration, admixture between populations, language, culture (e.g. the level of consanguineous marriages), and evolutionary processes like natural selection in the presence of factors such as climate change, changes in diet and exposure to infectious diseases (Campbell *et al.*, 2008, 2010; Uren *et al.*, 2016; Thami *et al.*, 2019; Chaichoompu *et al.*, 2020). Wright's fixation index, Fst, is the classic measure for comparing the genetic structure between different populations (Wright, 1949). It can range from zero to one with one indicating high genetic variation between populations and zero indicating low genetic variation between populations (Jakobsson *et al.*, 2013).

1.4 GLOBAL VARIATION OF ADME SNPs

Drug response and tolerance is known to vary not only between individuals but also among different ethnic groups, which lead to the development of the field of Pharmacoethnicity. Pharmacoethnicity has a fundamental impact on pharmacogenetic studies as it will help identify pharmacologically relevant markers that will help reduce ADRs and better understand the variation of drug response across different populations (O'Donnell *et al.*, 2009; Gamazon *et al.*, 2012; Ortega *et al.*, 2014). Since the allelic frequency of polymorphisms in pharmacogenes varies among different ethnic groups, it contributes significantly to inter-ethnic difference in drug response (Yasuda *et al.*, 2008; Rodrigues *et al.*, 2019).

The frequency of some pharmacogenomic markers covers nearly the entire frequency range. An example would be the SNP rs9923231 in *VKORC1*, which has been shown to influence the dosing of warfarin, an anticoagulant that prevents blood clots (Limdi *et al.*, 2008). Globally, the minor allele frequency (MAF) of this SNP ranges from 0.02 to 0.95. This results in warfarin having a wide range of dose requirements in different populations. Due to the allele frequency, many African-Americans require a high dose of warfarin, and in general Asians require lower doses, with and intermediate requirement in Caucasian populations (Dang *et al.*, 2005; Li *et al.*, 2014). Dosing should however, proceed on an individual level, guided by the presence or absence of the variant. As developing countries rely heavily on the US FDA or European Medical Agency (EMA) guidelines for guidelines on dosing, a comprehensive understanding of the inter-ethnic differences in ADME gene variation is critical to attain effective drug prescriptions globally (Phan *et al.*, 2009).

The frequency of certain other pharmacogenomic markers is similar across different population groups regardless of geography and ancestry. An example is the variant rs3918290 in the dihydropyrimidine dehydrogenase gene (*DPYD*). This variant is associated with ADRs like neutropenia, myelosuppression, increased risk of death, and increased severity of drug toxicity from capecitabine, a chemotherapeutic agent, which inhibits DNA synthesis in the targeted tumour when converted to 5-fluorouracil (Raida *et al.*, 2001; Salgueiro *et al.*, 2004; Largillier *et al.*, 2006; Schadt *et al.*, 2008; Schwab *et al.*, 2008; Sulzyc-Bielicka *et al.*, 2008).

The cytochrome P450 enzyme system consists of a family of isoenzymes predominantly active in the liver and the intestine. Each of these isoenzymes have their own substrate specificity, and catalyses oxidative reactions to protect the body from toxic substances by making them more soluble. With respect to drug metabolism, CYP3A4, CYP2C9, CYP2C19 and CYP2D6 are the most important cytochromes. Genes that code for these enzymes have several variant alleles that encode for enzymes with decreased activity. This gives rise to interindividual differences in the capacity to metabolise drugs (www.cypalleles.ki.se).

CYP gene variant allele distribution differs among different ethnic populations (Stubbins *et al.*, 1996; Zhou *et al.*, 2017). *CYP3A5* encodes for one among the four CYP3A subfamily of enzymes which is responsible for the metabolism of more than 50% of medicines prescribed worldwide (Zanger *et al.*, 2013). Enzymatic activity and expression levels of this gene are modulated by genetic polymorphisms. A 6986A>G transition within intron 3 (rs776746, *CYP3A5**3), leads to an incorrectly spliced mRNA and non-functional protein. This allele shows significant differences in frequency across different populations (Suarez-Kurtz *et al.*, 2014). Its allele frequencies range from 0.14 among sub-Saharan Africans to >0.95 in European populations (Kuehl *et al.*, 2001). In stark contrast to *CYP3A5*3*, *CYP3A5*6* (rs10264272, a 14690G>A transition that results in a mRNA splice variant) and *CYP3A5*7* (rs41303343, a 23132insT insertion that causes a premature stop codon), are two other defective CYP3A5 alleles that are relatively frequent in Africans but are rare or absent in Europeans (Hustert *et al.*, 2001).

1.5 GENETICALLY DETERMINED DIFFERENCES IN DRUG METABOLISM

Drug dosages are ascertained at a particular dose based on the assumption that the drug will yield a predetermined concentration in circulation that will bring about the desired therapeutic effect (Arbitrio *et al.*, 2016). Other factors that influence the relationship between drug dose and concentration in blood is how healthy the individual's liver and kidneys are and other medications that have been prescribed along with the drug under consideration. Environmental factors like alcohol consumption, diet, smoking, and caffeine intake will also alter the rate of drug metabolism (van Schaik, 2008). Differences in genetic makeup also affect

drug dose and its concentration in blood. In certain individuals, variant alleles of ADME/DMET genes that result in altered activity can give rise to fluctuations in circulating drug concentration in response to certain drug doses.

Individuals can be classified into four metaboliser types depending on their ability to metabolise drugs (Chang *et al.*, 2015):

1. Ultra-rapid metaboliser (UM): Ultra-rapid metabolisers have metabolising enzymes that are very active. Medications break down before they can have any effect and hence a higher dose of the drug might be required for the drug to work properly.
2. Normal/Extensive metaboliser (NM): These individuals are considered to have a normal rate of metabolism. They are likely to metabolise drugs normally. Medication is likely to work as intended at standard doses of the drug without any ADRs.
3. Intermediate metaboliser (IM): Have intermediate metabolising enzymes that are considered to be less active. Drugs are not metabolised as completely as a normal metaboliser. This means that the standard drug dose will not work and must be lowered. The lower drug dose will prevent the unmetabolised drug from building up in the body and causing ADRs.
4. Poor metaboliser (PM): Poor metabolisers experience a very slow breakdown of medications because the metabolising enzyme has very low activity. This makes side effects more pronounced even at a very low drug dose.

CYP2C9 is the main enzyme of the CYP2C subfamily of enzymes. It is a phase I drug-metabolising enzyme isoform that plays a major role in the oxidation of both xenobiotic and endogenous compounds. It is involved in the metabolism of 10-20% of most commonly prescribed drugs and is primarily expressed in the liver (Lee *et al.*, 2002; Ali *et al.*, 2009). Genetic variants in *CYP2C9* alter the metabolic activity of the encoded enzyme and result in ADRs. They are therefore a significant factor in pathogenesis (Van Booven *et al.*, 2010). The variant alleles identified for *CYP2C9* are listed in the online database PharmVar (https://www.pharmvar.org/gene/CYP2C9). The wild type allele *CYP2C9*1* has normal

enzyme activity and a high global allelic frequency of nearly 80%. *CYP2C9*2* (rs1799853,430C>T, Arg144Cys) and *CYP2C9*3* (rs1057910,1075A>C, Ile359Leu) have reduced enzyme activity. Both these variants have significantly lower frequencies in African and Asian populations compared with populations of European origin (van Schaik, 2008; Van Booven *et al.*, 2010). Poor enzyme activity in patients is associated with a higher risk of ADRs.

CYP2D6 accounts for approximately 1–4% of all hepatic CYP450 enzymes. It metabolises approximately 25% of the most commonly prescribed drugs (Ingelman-Sundberg *et al.*, 2007). Depending on the genotype of *CYP2D6*, which determines the inherent ability to metabolise CYP2D6 substrates, individuals can be categorised into four groups (Zanger *et al.*, 2004). In decreasing order of enzyme activity, they are also classified as UM, NM, IM, and, PM. PMs are mainly found in Europe, UMs are mainly found in North Africa and Oceania and IMs are predominantly found in Asia (Ingelman-Sundberg *et al.*, 2007). About 5–14% of Caucasians, 0–5% Africans, and 0–1% of Asians are known as poor metabolisers because they lack CYP2D6 activity (S. F. Zhou *et al.*, 2009).

CYP2C8 plays a role in the metabolism of paclitaxel, a chemotherapeutic drug. The variant alleles *CYP2C8*2*, **3*, **5*, **7* and **8* have decreased activity. *CYP2C8*2* is found predominantly in African-Americans (18%) but is less commonly found in Caucasians (allele frequency 0-0.7%). Patients with this haplotype have lower intrinsic clearance of paclitaxel. The *CYP2C8*3* allele is found in 2% of African-Americans and in 8-13% of Caucasians (van Schaik, 2008) .

1.6 FUNCTIONAL VARIANTS IN ADME GENES THAT AFFECT DRUG EFFICACY

In this section I focus on the following genes: *CYP2C19*, *CYP3A5*, *UGT1A1*, *DPYD*, *SULT1A1*, *CYP1A2* and *GSTM1* to understand in detail how variants in these ADME genes affect drug response and risk for ADRs. These genes have been extensively studied and therefore will help in understanding the importance and their clinical consequences.

1.6.1 CYP2C19

CYP2C19 is a phase I drug metabolising enzyme. The chromosomal location of the *CYP2C19* gene is 10q24.1-q24.3. It encodes for a 490-amino-acid protein, the enzyme cytochrome P450 2C19 (CYP2C19), which plays a major role in the metabolism of a wide array of therapeutic drugs, across different drug classes like (Li-Wan-Po *et al.*, 2010; Lee, 2013):

- Antidepressant drugs like Citalopram, Imipramine, Amitriptyline, Clomipramine
- Antiplatelet drugs like Clopidogrel
- Antiviral drugs like Nelfinavir
- Anticonvulsant drugs like Mephenytoin
- Antacid drugs like Omeprazole
- Antifungal drugs like Voriconazole
- Antimalarial drugs like Proguanil
- Anti-oestrogen drugs like Tamoxifen
- Anxiolytic drugs like diazepam
- Beta blocker drugs like Propranolol
- Antiulcer drugs like Omeprazole
- Cytotoxic agents like Cyclophosphamide, Teniposide

The *CYP2C19* gene is highly polymorphic. All the known variants of *CYP2C19* can be accessed at https://www.pharmvar.org/gene/CYP2C19. *CYP2C19*1* is the wild type allele that is associated with functional CYP2C19-mediated metabolism (Hulot *et al.*, 2006). Polymorphisms in *CYP2C19* are associated with isozymes with reduced enzyme activity. *CYP2C19 *2*, **3*, **4*, **5*, **6*, **7*, **8*, **16* and **26* alleles are examples of loss-of-function alleles that result in isozymes with reduced enzyme activity (Scott *et al.*, 2013; Zhu *et al.*, 2016). *CYP2C19*2* is the most common loss-of-function allele with allele frequencies of ~15% in Caucasians and Africans, and 29–35% in Asians (Scott *et al.*, 2013; Zhu *et al.*, 2016). With the exception of *CYP2C19*3*, which is found in Asians with an allele frequency of 2–9%, other CYP2C19 variant alleles with reduced or absent enzyme activity are present in much lower frequencies, lower than 1%, in all populations (Scott *et al.*, 2013; Zhu *et al.*, 2016).

*CYP2C19*17* allele (c.-806C>T; rs12248560), however, results in an isozyme with increased enzyme activity (Sim *et al.*, 2006).

Based on which *CYP2C19* genotype an individual has, they can be categorised into ultra-rapid metaboliser, extensive metaboliser, intermediate metaboliser, and poor metabolisers of drugs. This is further elaborated in Table 1.1. 12-23% of Asians, 3-5% of Caucasians and 4% of Africans populations are CYP2C19 poor metabolisers (Bathum *et al.*, 1999; Dandara *et al.*, 2001; van Schaik, 2008). Since medications metabolised by *CYP2C19* will undergo very slow breakdown because of very low enzyme activity, the side effects of these medications even at very low doses will be pronounced.

Table 1.1: CYP2C19 metabolism phenotypes based on genotype (Hicks *et al.*, 2017)

Likely phenotype	**Genotypes**	**Examples of diplotypes**
Ultra-rapid metaboliser	An individual carrying two increased activity alleles (*17) or one functional allele (*1) and one increased activity allele (*17)	**17/*17, *1/*17*
Extensive metaboliser	An individual carrying two functional alleles (*1)	**1/*1*
Intermediate metaboliser	An individual carrying a functional allele (*1) and a loss-of-function allele (*2-*8) or an increased activity allele (*17) and a loss-of-function allele (*2-*8)	**1/*2, *1/*3, *2/*17*
Poor metaboliser	An individual carrying two loss-of-function alleles (*2-*8)	**2/*6, *3/*5, *7/*7*

For example, after administration of diazepam, a drug used to treat anxiety disorders, alcohol withdrawal symptoms, and muscle spasms, prolonged sedation and unconsciousness has been observed in CYP2C19 PMs (Bertilsson, 1995; Jose *et al.*, 2016). CYP2C19 PMs and CYP2C19 IMs respond poorly to clopidogrel and show reduced platelet inhibition compared

to CYP2C19 NMs (Bura *et al.*, 2006; Marian, 2009). An alternate drug must be considered in CYP2C19 PMs as they may not benefit from clopidogrel treatment. Though patients with a *CYP2C19*17* allele respond better to clopidogrel, they show a higher risk for bleeding (Sibbing *et al.*, 2010). Clopidogrel is found to exert a protective effect on *CYP2C19*17* allele carriers after myocardial infarctions (Tiroch *et al.*, 2010). CYP2C19 PMs respond more to omeprazole and lansoprazole for gastric ulcers caused by *Helicobacter pylori* than in CYP2C19 NMs due to higher concentrations of the drugs in PMs (Sohn *et al.*, 1997; Furuta *et al.*, 2004).

In general, following *CYP2C19* genotyping, clinical decision strategies suggest two regimens: adjusting the drug dose according to the underlying genotype or prescribing an altogether different drug. The greatest interindividual variation is observed in the IM group and hence this group is the greatest uncertainty when it comes to deciding the dosage of the drug. The underlying mechanism for this high variation remains unclear. Integration of clinical, environmental and drug-response modulating factors will help in the better understanding of this variation.

1.6.2 CYP3A5

The cytochrome P450 (CYP) 3A subfamily participates in the biotransformation of approximately 50% of drugs known to undergo oxidative metabolism (Evans *et al.*, 1999). In humans, four members of the CYP3A subfamily have been identified. They are: *CYP3A4*, *CYP3A5*, *CYP3A7*, and *CYP3A43*. Of these, *CYP3A5* is the second most abundant (Park *et al.*, 2014). *CYP3A4* and *CYP3A5* account for ~30% of hepatic cytochrome P450 (Lamba *et al.*, 2012). The genes coding for these enzymes are located in a 231-kb cluster on chromosome 7q21-7q22 adjacent to each other in the order *CYP3A5*, *CYP3A7*, *CYP3A4* and *CYP3A43*. *CYP3A5* is a 31.8 kb segment (Hustert *et al.*, 2001). The gene is on the minus chromosomal strand. It has 9 exons and expresses a 502 amino-acid protein (Lamba *et al.*, 2012).

CYP3A5 is highly polymorphic. All *CYP3A5* allelic variants are listed on the Pharmacogene Variation Consortium website (https://www.pharmvar.org/gene/CYP3A5). *CYP3A5**1, the

wild type allele, codes for the functional CYP3A5. Individuals carrying at least one wild type allele express high levels of CYP3A5 enzyme (Hustert *et al.*, 2001; Kuehl *et al.*, 2001). The most common and non-functional variant of *CYP3A5* is *CYP3A5*3* (rs776746). *CYP3A5*3* allele causes alternative splicing and results in lost CYP3A5 enzyme activity (Kuehl *et al.*, 2001; Herbert *et al.*, 2005). Allele frequencies of *CYP3A5* vary widely across different population groups. The estimated allele frequency of *CYP3A5*3* in populations of European ancestry is 0.82–0.95 (Kuehl *et al.*, 2001). *CYP3A5*3* allele frequencies in different populations are 0.33 in African-Americans; 0.85 in the Japanese; 0.65 in the Chinese; 0.75 in Mexicans; 0.67 in Southeast Asians excluding Japanese and Chinese; 0.65 in Pacific islanders; 0.06 in Nigerians, 0.96 in Basques and 0.4 in Southwest American Indians (Kuehl *et al.*, 2001).

*CYP3A5*2* (rs28365083, g.27289C > A, T398N) is not a fully functional allele (van Schaik *et al.*, 2002). This variant has been found in different populations of European ancestry (van Schaik *et al.*, 2002; Petrova *et al.*, 2007; Ekhart, Doodeman, *et al.*, 2009). It has not been found when assayed for in other populations (Hiratsuka *et al.*, 2002; Shih *et al.*, 2002; Dandara *et al.*, 2005; Krishnakumar *et al.*, 2012). The *CYP3A5 *5* allele results in the complete absence of enzyme activity (Fröhlich *et al.*, 2004; Birdwell *et al.*, 2015).

*CYP3A5*6* (rs10264272; 14690G > A) is another non-functional allele. It is a guanine to adenine transition in exon 7 that results in exon 7 skipping (alternative splicing) (Kuehl *et al.*, 2001). This causes protein truncation and culminates in the absence of CYP3A5 protein. It is predominantly present in African-Americans (5-25%) and is found to be very rare in European and Asian populations (Kuehl *et al.*, 2001; Balram *et al.*, 2003; Lee *et al.*, 2003).

*CYP3A5*7* (rs76293380, 27131–27132ins T) is an insertion between codon 345 and 346 that results in a shift in reading frame (Hustert *et al.*, 2001). This shift results in a truncated and non-functional protein. *CYP3A5*7* is found in the African population at a frequency of 8% (Lee *et al.*, 2003) but has not been found in European (van Schaik *et al.*, 2002; Lee *et al.*, 2003) or Asian populations (Lee *et al.*, 2003; Park *et al.*, 2008).

Depending on which *CYP3A5* genotype people carry, they can be classified as extensive metaboliser, intermediate metaboliser or poor metaboliser of drugs. This is further elaborated in Table 1.2.

Table 1.2: CYP3A5 metabolism phenotypes based on genotype (Birdwell *et al.*, 2015)

Likely phenotype	**Genotypes**	**Examples of diplotypes**
Extensive/Normal metaboliser	An individual carrying two functional alleles	**1/*1*
Intermediate metaboliser	An individual carrying one functional allele and one non-functional allele	**1/*3*, **1/*6*, **1/*7*
Poor metaboliser	An individual carrying two non-functional alleles	**3/*3*, **6/*6*, **7/*7*, **3/*6*, **3/*7*, **6/*7*

*CYP3A5*2*, **8*, and **9* are rare variants of unknown functional significance. If a copy of *CYP3A5 *1* is present, the phenotype is expected to be intermediate metaboliser.

CYP3A5 extensive and intermediate metabolisers metabolise CYP3A substrates more rapidly than CYP3A5 poor metabolisers. For both extensive and intermediate metabolisers, the dose of the CYP3A5 substrate medication will/might have to be increased. Approximately 20% of African-Americans are CYP3A5 normal metabolisers (Roy *et al.*, 2005). Less than 1% of people of European descent are CYP3A5 normal metabolisers (Suarez-Kurtz *et al.*, 2014). About 50% of African-Americans are CYP3A5 intermediate metabolisers, while only 15% of people of European descent are CYP3A5 intermediate metabolisers (Roy *et al.*, 2005). 30% of African-Americans are CYP3A5 poor metabolisers while 85% of people of European descent are CYP3A5 poor metabolisers (Bains, 2013; Suarez-Kurtz *et al.*, 2014; Birdwell *et al.*, 2015).

Tacrolimus is prescribed to prevent post-transplantation organ rejection. Individuals homozygous for *CYP3A5*1* have a higher tacrolimus clearance rate than individuals heterozygous for *CYP3A5*1*, who in turn have a higher tacrolimus clearance rate than individuals homozygous for *CYP3A5*3/*3* (Passey *et al.*, 2011). Ideally, the concentration of

tacrolimus must be high enough to prevent transplant rejection (O'Seaghdha *et al.*, 2009), but low enough to not cause toxicity (Borobia *et al.*, 2009). Hence, the trough concentration (trough concentration/trough level (C_{trough})) is the lowest concentration reached by a drug before the next dose is administered. The trough concentration of Tacrolimus is diligently monitored post transplantation and dosage level is adjusted accordingly. Thus there is an interdependence between *CYP3A5* genotype and tacrolimus dosage levels and this association plays a significant role in preventing acute organ rejection (MacPhee *et al.*, 2004; Renders *et al.*, 2007; Quteineh *et al.*, 2008). The dosing equation for tacrolimus considers five factors. They are *CYP3A5* genotype, number of days after transplant, age of the person, whether transplant was done at a steroid sparing centre or not and use of calcium channel blockers. Hence depending on the factors involved in the dosing equation, tacrolimus dosing and drug concentration will change.

The genotype *CYP3A5*1/*3* is associated with rapid clearance of the antiretroviral drug saquinavir compared to *CYP3A5*3/*3* (Fröhlich *et al.*, 2004; Mouly *et al.*, 2005). *CYP3A5* expressers have been found to have less treatment-related neurotoxicity in paediatric patients with precursor B-cell acute lymphoblastic leukaemia (ALL) treated with vincristine (Egbelakin *et al.*, 2011). Since *CYP3A5*1* is linked to toxicity in advanced renal-cell carcinoma patients treated with sunitinib, it is also associated with subsequent dose reductions in these patients (Garcia-Donas *et al.*, 2011). *CYP3A5* expressors have also been found to have higher metabolism rates for cyclosporine A (an immunosuppressant) (Dai *et al.*, 2004) and ifosfamide (a chemotherapy drug) (McCune *et al.*, 2005).

CYP3A5 has also been shown to be a determinant in the how effectively statins lower lipid levels. In a study, individuals with the alleles **1* than **3* were found to respond poorly to the statin drugs lovastatin, simvastatin, and atorvastatin (Kivistö *et al.*, 2004; Willrich *et al.*, 2008). The *CYP3A5 *3/*3* genotype is also found to be associated with more severe ADRs to statin treatment than individuals with other *CYP3A5* genotype (Wilke *et al.*, 2005).

CYP3A5 has been found to be expressed more frequently in African-Americans than in Caucasians (Hustert *et al.*, 2001; Daly, 2006). CYP3A5 makes up 50% of the total hepatic CYP3A enzymes in people. This means that *CYP3A5*'s contribution to inter-individual and inter-ethnic differences in CYP3A-dependent drug response and clearance is quite significant (Kuehl *et al.*, 2001).

1.6.3 UGT1A1

Uridine diphosphate glucuronosyl transferase1A1 (*UGT1A1*) is a phase II drug metabolising enzyme (Schulz *et al.*, 2009; Barbarino *et al.*, 2014). It plays a major role in the conjugation and subsequent elimination of potentially toxic xenobiotics, carcinogens and pharmaceutical drugs. It catalyses the glucuronidation of a large variety of compounds like oestrogens, bilirubin and the chemotherapeutic drug irinotecan (Schulz *et al.*, 2009).

The *UGT1A1* gene is located at 2q37and extends about 160 kb. Variants in the promoter and coding regions of the gene are found to affect the activity of the enzyme. The different alleles of *UGT1A1* and the functional status associated with each allele is listed in Table 1.3.

Table 1.3: Functional status of alleles of *UGT1A1* (Gammal *et al.*, 2016)

Allele	Allele Functional Status
**1*	Normal function
**6*	Decreased Function
**27*	Decreased Function
**28*	Decreased Function
**36*	Increased Function
**37*	Decreased Function
**60*	Normal function
**80*	Uncertain Function

*UGT1A1*80* is in very high LD with **28* and **37*. When in LD with **28* or **37*, UGT1A1 has decreased function. The function of the **80* allele alone, in the absence of **28* or **37*, is uncertain. If a patient is only tested for **80*, decreased function may be inferred.

The most frequent genetic variant that affects UGT1A1 function is a dinucleotide TA_n repeat polymorphism (rs8175347) located in a TATAA consensus element in the *UGT1A1* promoter at –53 relative to the translation start site. This varies from five to eight TA repeats (Palomaki *et al.*, 2009). *UGT1A1*1* (TA_6) is the fully functional allele with normal enzyme activity. Depending on the genotype of *UGT1A1*, individuals can be classified as extensive metaboliser, intermediate metaboliser or poor metaboliser of drugs (Gammal *et al.*, 2016). This is further elaborated in Table 1.4.

Table 1.4: UGT1A1 metabolism phenotypes based on genotype (Gammal *et al.*, 2016)

Likely phenotype	**Genotypes**	**Examples of diplotypes**
Extensive metaboliser	An individual carrying two functional alleles (*1) and/or increased function alleles (*36). Alternatively identified by homozygosity for rs887829 C/C.	**1/*1; *1/*36; *36/*36; rs887829 C/C*
Intermediate metaboliser	An individual carrying one functional allele (*1) or increased function allele (*36) plus one decreased function allele (*6, *28, *37). Alternatively identified by heterozygosity for rs887829 C/T.	**1/*28; *1/*37; *36/*28; *36/*37; rs887829 C/T, *1/*6*
Poor metaboliser	An individual carrying two decreased function alleles (*6, *28, *37). Alternatively identified by homozygosity for rs887829 T/T (*80/*80)	**28/*28; *28/*37; *37/*37; rs887829 T/T (*80/*80), *6/*6*

UGT1A1 is the sole enzyme responsible for the metabolism of bilirubin in the liver (Strassburg, 2008). Irinotecan, an anti-cancer drug, is a topoisomerase I inhibitor. It is primarily used to treat metastatic and recurrent colorectal cancer. It is a prodrug that is converted to SN-38, a metabolite that is highly toxic because it has 100-fold higher anti-tumour activity than its parent compound (Ramesh *et al.*, 2010). *UGT1A1* is predominantly responsible for the glucuronidation of SN-38. Irinotecan has a very narrow therapeutic range. The most common

dose-limiting adverse effects of irinotecan, neutropenia, diarrhoea, and asthenia are believed to be caused by higher or more prolonged exposure to the active form SN-38. These side effects can be severe enough to either reduce the dosage of the drug or discontinue the drug (Lankisch *et al.*, 2008; Barbarino *et al.*, 2014). *UGT1A1*28* and **6* alleles are most associated with the development of irinotecan toxicities. Approximately 7% of patients treated with irinotecan treatment who have severe neutropenia and fever die from the complications (Obradovic *et al.*, 2008).

Individuals homozygous for the fully functional **1* alleles have average levels of the enzyme and will metabolise SN-38 more quickly than those who are either heterozygous or homozygous for alleles that confer decreased enzyme activity. If irinotecan dosage can be adjusted according to *UGT1A1* genotype a significant proportion of ADRs can be avoided. Hence, the extent of side effects of irinotecan treatment can be predicted by variations in *UGT1A1* (Palomaki *et al.*, 2009). *UGT1A1* genotyping before irinotecan treatment therefore can be recommended in clinical practice because it has been found to be cost-effective (Gold *et al.*, 2009).

*UGT1A1*6/*6* is associated with lower clearance of etoposide (an anticancer agent) in people of African ancestry (Kishi *et al.*, 2004). *UGT1A1*28/*28* is associated with slower elimination of the drug raloxifene (prescribed for postmenopausal women with osteoporosis to reduce the risk of invasive breast cancer and in postmenopausal women at high risk of invasive breast cancer) (Palomaki *et al.*, 2009; Trontelj *et al.*, 2009). Hyperbilirubinemia is linked to the **28* allele in patients treated with atazanavir (an antiretroviral medication) (Lankisch *et al.*, 2006) and tranilast (an antiallergic medication) (Danoff *et al.*, 2004) .

*UGT1A1*6* is a widespread diplotype in individuals of East Asian ancestry (Korean, Japanese, Chinese). This diplotype is all but is absent in European and African populations (Dai *et al.*, 2013; Liu *et al.*, 2007; Palomaki *et al.*, 2009). Therefore, in the Asian population, patients heterozygous or homozygous for the **6* allele have a higher risk for neutropenia, diarrhoea

and increased exposure to the cytotoxic SN-38 metabolite (Onoue *et al.*, 2009). Therefore, this allele can be potentially used for the prediction of irinotecan-induced toxicities in Asians (Barbarino *et al.*, 2014).

The *UGT1A1*36* allele which is associated with increased enzymatic activity is found in higher frequencies in African and African-American populations (Beutler *et al.*, 1998; Palomaki *et al.*, 2009). Increased enzymatic activity will result in medications metabolised by UGT1A1 breaking down before they can exert any effect. Therefore, in people homozygous for this allele, drugs will have to be administered in higher doses to attain optimal drug efficacy.

1.6.4 DPYD

The *DPYD* gene is an important pharmacogene that codes for the primary enzyme, dihydropyrimidine dehydrogenase (DPD), which is involved in breaking down fluoropyrimidine to its inactive metabolites. It spans about 950kb on chromosome 1p22. DPD catalyses "the first and the rate-limiting step in the breakdown of the pyrimidine nucleotides thymine and uracil" (Dean, 2016). This enzyme also catalyses the rate-limiting step in fluorouracil metabolism, an anticancer drug belonging to the fluoropyrimidines group of drugs. It catalyses 85% of administered fluoropyrimidine (Boige *et al.*, 2016). "Fluorouracil is used in the palliative management of carcinomas of the colon, rectum, breast, stomach, and pancreas" (Amstutz *et al.*, 2009). Treatment with fluoropyrimidines can cause severe to potentially life-threatening toxicity. Approximately 10-40% of patients develop ADRs (Amstutz *et al.*, 2009). Occurrence and severity of ADRs in patients who have been administered fluoropyrimidine are partly explained by clinical factors like age and sex (Boige *et al.*, 2016). Severe ADRs typically lead to the "discontinuation of effective anticancer therapy, and often requires hospitalization" (Deenen *et al.*, 2015).

The *DPYD* gene is a relatively large gene and given that the variants occur at low frequencies, only some of them have an assigned star (*) allele. Therefore, unlike the star allele nomenclature that is used for other DMEs, the nomenclature used for *DPYD* variants is rs IDs.

(Amstutz *et al.*, 2018). Polymorphisms in the *DPYD* gene cause DPD deficiency. The CPIC lists 82 DPYD variants (Amstutz *et al.*, 2018). Table 1.5 lists the known alleles of *DPYD* and their associated functions.

Table 1.5: Functional status of alleles of *DPYD* (Amstutz *et al.*, 2018)

Allele	Allele Functional Status
c.1905+1G>A (*2A)	No function
c.1898delC (*3)	No function
c.1601G>A (*4)	Normal function
c.1627A>G (*5)	Normal function
c.2194G>A (*6)	Normal function
c.295_298delTCAT (*7)	No function
c.703C>T (*8)	No function
c.85T>C (*9A)	Normal function
c.2657G>A (*9B)	Normal function
c.2983G>T (*10)	No function
c.1003G>T (*11)	Normal function
c.1156G>T (*12)	No function
c.1679T>G (*13)	No function
c.1129-5923C>G, c.1236G>A (HapB3)	Decreased function
c.2846A>T	Decreased function
c.557A>G	Decreased function
c.62G>A	Normal function
c.496A>G	Normal function
c.1218G>A	Normal function
c.1896T>C	Normal function
c.46C>G	Normal function
c.61C>T	No function
c.313G>A	Normal function
c.343A>G	Normal function
c.451A>G	Normal function
c.498G>A	Normal function
c.601A>C	No function
c.632A>G	No function
c.775A>G	Normal function
c.868A>G	Decreased function
c.929T>C	Normal function
c.934C>T	Normal function
c.967G>A	Normal function

c.1024G>A	No function
c.1057C>T	No function
c.1108A>G	Normal function
c.1181G>T	Normal function
c.1180C>T	Normal function
c.1260T>A	Normal function
c.1278G>T	Normal function
c.1294G>A	Normal function
c.1314T>G	Decreased function
c.1349C>T	Normal function
c.1358C>G	Normal function
c.1403C>A	Normal function
c.1475C>T	No function
c.1484A>G	No function
c.1519G>A	Normal function
c.1543G>A	Normal function
c.1577C>G	Normal function
c.1615G>A	Normal function
c.1682G>T	Normal function
c.1775G>A	No function
c.1774C>T	No function
c.1777G>A	No function
c.1796T>C	Normal function
c.1905C>G	Normal function
c.1906A>C	Normal function
c.1990G>T	Normal function
c.2021G>A	No function
c.2161G>A	Normal function
c.2186C>T	Normal function
c.2195T>G	Normal function
c.2279C>T	Decreased function
c.2303C>A	Normal function
c.2336C>A	Normal function
c.2482G>A	Normal function
c.2582A>G	Normal function
c.2623A>C	Normal function
c.2639G>T	No function
c.2656C>T	Normal function
c.2872A>G	No function
c.2915A>G	Normal function
c.2921A>T	Normal function
c.2933A>G	No function
c.2978T>G	Normal function

c.2977C>T	Normal function
c.3049G>A	Normal function
c.3061G>C	Normal function
c.3067C>A	Normal function
c.525G>A	Normal function
c.1371C>T	Normal function

Depending on the genotype of *DPYD,* individuals can be categorised into three broad categories as shown in Table 1.6 (Henricks *et al.*, 2015).

Table 1.6: DPYD metabolism phenotypes based on genotype (Henricks *et al.*, 2015)

Likely Phenotype	**Genotypes**	**Examples of genotypes**
Normal Metaboliser	An individual carrying two normal function alleles	*4/*4, *4/*5, *11/c.[62G>A]
Intermediate Metaboliser	An individual carrying two decreased function alleles or and individual carrying one normal function allele plus one non-functional allele or one decreased function allele.	*9B/*13, *11/c.[557A>G], c.[2846A>T]/c.[2846A>T]
Poor Metaboliser	An individual carrying two non-function alleles or an individual carrying one non-functional plus one decreased function	*12/13, c.[61C>T]/c.[868A>G]

The risk of severe fluoropyrimidine toxicity increases significantly with the presence of the *DPYD* variants that are associated with low DPD activity (Caudle *et al.*, 2013). Therefore, patients may benefit from being categorised on the basis of the *DPYD* genotype (Dean, 2016). The most clinically relevant polymorphism in the *DPYD* gene is *DPYD*2A* (c.1905+1G>A; IVS14+1G>A; rs3918290). As shown in Table 1.5, *DPYD*2A* results in a truncated protein without any residual enzyme activity. *DPYD*2A* allele frequency has been reported to vary between 0.1% to 1.0% in African–American and Caucasian populations respectively (Caudle *et al.*, 2013; Offer *et al.*, 2014; Henricks *et al.*, 2015). The allele frequency of *DPYD*13*, another

DPYD variant that results in no enzyme activity (c.1679T>G; I560S; rs55886062) was found to vary from 0.07 to 0.1% in Caucasians (Caudle *et al.*, 2013; Lee *et al.*, 2014).

Patients treated with fluoropyrimidine drugs are at a higher risk of drug toxicity (possibly to even fatal levels) if they are poor or intermediate DPYD metabolisers. It is recommended that 5-fluorouracil or 5-fluorouracil prodrug-based regimens must be avoided in poor metabolisers. An alternate drug should be recommended. If alternative drugs are not considered a suitable therapeutic option for poor metabolisers, 5-fluorouracil may only be prescribed at a significantly reduced dose with frequent monitoring of the patient (Caudle *et al.*, 2013; Dean, 2016; Amstutz *et al.*, 2018).

1.6.5 SULT1A1

The *SULT1A1* gene resides on chromosome 16. It has 7 coding exons and 2 alternatively spliced upstream non-coding exons. The gene encodes a 295 amino acid protein, the sulphotransferase 1A1 (SULT1A1) (Hildebrandt *et al.*, 2009). Human cytosolic sulfotransferases (SULTs) are members of the SULT gene superfamily involved in Phase II biotransformation reactions (Jakoby *et al.*, 1990). Sulphotransferase 1A1 is the most widely expressed member of the sulphotransferase family. The enzyme is present in many human tissues, including liver, kidney and platelets (Hildebrandt *et al.*, 2009).

*SULT1A1*1* is the wild type allele. *SULT1A1*2* (rs9282861) is a non-synonymous G to A transition at nucleotide 638 in exon 7 that results in an amino acid change at codon 213 from arginine to histidine. The *SULT1A1*2* allele confers lower enzymatic activity (Raftogianis *et al.*, 1997). Its allele frequencies in Caucasians, Chinese and African-Americans are 33%, 8% and 29% respectively (Carlini *et al.*, 2001).

*SULT1A1*3* (rs1801030) is an A to G conversion at nucleotide 667 that results in a nonsynonymous change in codon 223 from methionine to valine (p.M223V). This variant is

more prevalent in African–Americans than in Caucasians and Chinese and confers lower enzyme activity. Its allele frequency is 23%, 1% and 0.6%, respectively (Carlini *et al.*, 2001).

rs3760091(-624G>C) is present in the 5' flanking region of *SULT1A1*. It is present at allele frequencies greater than 30% for Caucasians, African-Americans and the Chinese. rs3760091 shows high LD with *SULT1A1*2*. The C allele of this variant is associated with increased SULT1A1 activity (Ning *et al.*, 2005; Hildebrandt *et al.*, 2009).

rs750155 (-396G>A) is present in the promoter region of *SULT1A1* (Ning *et al.*, 2005; Hildebrandt *et al.*, 2009). This variant is in LD with *SULT1A1*2*. The allele frequencies of this variant are 48%, 63% and 42% in Caucasian, African-American and Chinese populations respectively (Ning *et al.*, 2005). The C allele is associated with decreased SULT1A1 activity in African-Americans and increased activity in Caucasians (Ning *et al.*, 2005; Hildebrandt *et al.*, 2009).

1.6.6 CYP1A2

Cytochrome P450 family 1 subfamily A member 2 (CYP1A2) is a phase 1 DME. The gene encoding CYP1A2 has six exons, extends 7.8 kb and is located at 15q22 (Ikeya *et al.*, 1989). It has more than 150 variants (http://www.cypalleles.ki.se). CYP1A2 is one of the major CYPs in the human liver. It accounts for ∼13–15% of all CYPs (Shimada *et al.*, 1994). This enzyme is involved in the metabolism of several chemicals and drugs like caffeine, clozapine, olanzapine, amitriptyline, R-warfarin, veramapil, acetaminophen, clomipramine, theophylline, haloperidol and imipramine There is large interindividual variability in the elimination of drugs that are metabolised by CYP1A2. This variability has been attributed to genetic mutations and environmental factors (Thorn *et al.*, 2012).

*CYP1A2*1A* is the wild type or reference allele. The most extensively studied polymorphisms of *CYP1A2* are *CYP1A2*1C* (−3860G>A; rs2069514), *CYP1A2*1D* (−2467delT), CYP1A2**1E* (−739T>G; rs2069526), and CYP1A2**1F* (−163C>A; rs762551), as well as an intronic variant

and *CYP1A2*1K* haplotype that is defined by the three variants rs2069526, rs12720461 (-729C>T), and rs762551 (Thorn *et al.*, 2012).

There is an association between C*YP1A2*1F* and high risk for leflunomide-induced host toxicity (S.-F. Zhou *et al.*, 2009). Leflunomide is an antirheumatic drug. CYP1A2**1F* is found to be associated with caffeine metabolism in Swedish smokers, with the A allele of CYP1A2**1F* being associated with increased metabolism of caffeine (Ghotbi *et al.*, 2007). CYP1A2**1F*/ rs762551 is also found to be associated with response to antipsychotic drugs and ADRs in individuals treated with them. (Özdemir *et al.*, 2001; Eap *et al.*, 2004; Tay *et al.*, 2007; Laika *et al.*, 2010; Lin *et al.*, 2010).

CYP1A2 activity has been observed to be markedly different across different ethnicities. Swedes have been found to have a 1.54-fold higher CYP1A2 activity than Koreans (Kall *et al.*, 1995). Asians and Africans have lower CYP1A2 activity than Caucasians (Relling *et al.*, 1992; S.-F. Zhou *et al.*, 2009). The frequency of the most well studied SNP in *CYP1A2*, CYP1A2**1F*/ rs762551 is found to vary widely across populations. The C allele frequencies range from 30-39% in Asians, 40-51% in Africans and 29-33% in Europeans (Gunes *et al.*, 2008).

1.6.7 GSTM1

Glutathione S-transferase enzyme (GSTM1) is another predominant phase II drug metabolising enzyme. The gene encoding GSTM1 is located at 1p13.3. The enzyme plays a significant role in the biotransformation and detoxification of xenobiotics and carcinogens (halomethanes and methyl bromide) (Chenevix-Trench *et al.*, 1995). Genetic polymorphisms and changes in the expression of *GSTM1* are thought to have a major impact on cancer susceptibility, inter-individual variability in prognosis, drug effects and toxicity (Kalacas *et al.*, 2019; Chanhom *et al.*, 2020; Tangkhuenkhan *et al.*, 2020). *GSTM1* is highly polymorphic. Individuals who carry the null genotype referred to as *GSTM1*0/*0* due to homozygous gene deletion, do not have the enzyme. The null genotype of *GSTM1* gene is shown to increase the risk of bladder, gastric, colon and lung cancers (You *et al.*, 2016). Glutathione S-transferase

genetic polymorphisms have been found to be associated with treatment outcomes of paclitaxel- and cisplatin-based chemotherapy (Medeiros *et al.*, 2003). The overexpression of the enzyme, glutathione S-transferase, is linked with resistance to chemotherapy (Huang *et al.*, 2008). Therefore, *GSTM1* is a gene that determines disease susceptibility and is also pharmacogenetically relevant.

1.7 PHARMACOGENOMICS AND PRECISION MEDICINE, A HISTORIC OVERVIEW

The Human Genome Project declared that the concept of personalized medicine was something to which they aspired (Schwab *et al.*, 2012b). Personalised medicine is a shift in approach from 'one-size-fits-all' to tailored care and targeted therapies based on an individual's genotype and the unique environmental factors that affect them (Brittain *et al.*, 2017; Carrasco-Ramiro *et al.*, 2017). The explosion of data generated by new omics technologies on the genetic, environmental, and phenotypic basis of human disease and drug response have emphasized the uniqueness of each individual. The application of these technologies will significantly enhance our understanding of the underlying molecular mechanisms of disease and improve our capacity for personalised interventions.

Understanding the role DNA plays in health, can and will transform healthcare practice from managing an illness to moving towards promoting health by means of the four Ps (Carrasco-Ramiro *et al.*, 2017):

- **P**rediction and Prevention of disease
- More **P**recise diagnoses
- **P**ersonalised and targeted interventions
- **P**articipatory role for patients

Personalized medicine or precision medicine is not a new concept. The historical writings of Hippocrates, Garrod, Osler, and others recognized the variability of treatment response and human health. The knowledge that genetic differences have the capability "to affect response to a biochemical environment has been known for over 100 years" (Issa, 2007). This

knowledge of genetic factors that influence drug efficacy and safety is of major relevance for personalized therapy (Arbitrio *et al.*, 2016). One of the first indications of this was based on studies on alkaptonuria early in the 20th Century (Garrod, 1902). In those studies, it was found that the enzyme, homogentisate oxidase, did not properly oxidise homogentisic acid, resulting in its accumulation. Alkaptonuria has the potential side effect of arthritis. Another observation was made when African-American soldiers took the antimalarial drug primaquine. It was noticed that some of them suffered from haemolysis, a reaction that was not common among European-Americans. This was a prime example for how genetic variation elicited a difference in drug response (Evans, Manley and McKusick, 1960; Mason and Russell, 1971; Bright *et al.*, 2013). It was later found that primaquine-induced haemolysis was caused by a gene variant that led to glucose-6-phosphate dehydrogenase deficiency (Carson *et al.*, 1956). *CYP2D6*4* was then found to be associated with primaquine treatment failure (Bennett *et al.*, 2013).

The impact that pharmacogenetics could have in the clinical care of patients was highlighted by Kalow and Gunn (1957) when they noticed longer-lasting "neuromuscular blockade and apnea" in some patients undergoing electroconvulsive therapy with succinylcholine, a neuromuscular blocking agent (Kalow *et al.*, 1957). The knowledge of a person's genotype has therefore time and again proven to be beneficial in the treatment of individuals. All these observations laid an early foundation for the field of pharmacogenetics and precision medicine.

1.8 IMPORTANCE OF PHARMACOGENETIC TESTS AND PERSONALIZED THERAPY IN PHARMACOGENOMICS

Pharmacogenomics promotes the development of targeted therapies that will best maximise drug efficacy and minimise ADRs by using a patient's genetic data and phenotypic information (Zanger *et al.*, 2008). A targeted therapy that got approved by the US FDA and European Medicines Agency after much pharmacogenomic research is the drug ivacaftor (Schwab *et al.*, 2012a). Ivacaftor is used for the treatment of a small subset of cystic fibrosis patients who have the *G551D* genetic variant in the cystic fibrosis transmembrane regulator (*CFTR*) gene.

Ivacaftor is approved only for cystic fibrosis patients bearing the specific *G551D* genetic variant, that impairs normal channel gating activity that regulates chloride and water transport across the cell membrane of epithelia cells. Ivacaftor targets the CFTR protein and increases its activity thereby improving lung function (Eckford *et al.*, 2012).

The goal of personalised therapy is to provide treatment tailored to the specific needs of an individual and to predict the clinical outcome of different treatments in different patients. Its potential applications in pharmacogenomics are (Kurth, 2000; Issa, 2002):

i. Determine how safe and effective an experimental new drug is prior to approval;
ii. Identify an individual's genetic susceptibility to diseases/genetic conditions;
iii. To protect people from undesired adverse drug reactions right from the beginning of a new drug therapy by identifying poor metabolisers.

The standard dosage of a drug may not be equally effective in all patients; for some it might be toxic, in others highly effective and not effective in another group (Umamaheswaran *et al.*, 2014). This wide variability in drug responses in individuals to standard doses of drugs is an issue when it leads to ADRs or therapeutic failures. By identifying polymorphisms in ADME genes that are involved in the various stages of drug metabolism and non-genetic factors that contribute to an individual's capacity to metabolise drugs like age, gender, nutrition, concomitant medications, organ function, co-morbidity and drug interactions, the number of patients responding to the drug (drug efficacy) can be increased and the number of patients affected by ADRs (drug toxicity) can be reduced significantly (Chowbay *et al.*, 2005; Zanger *et al.*, 2008; Sissung *et al.*, 2012; Sugarman *et al.*, 2016). Pharmacogenetics thus is a valuable tool that will allow clinicians to deliver a smarter, more fine-tuned approach to treating patients by predicting an individual's response to drug and propensity for drug toxicity. It will help clinicians arrive at an informed decision on optimal drug dosage guidance, based on that person's unique genetic makeup (Umamaheswaran, Kumar and Adithan, 2014; Saldivar *et al.*, 2016).

A pharmacogenetic test is a precision medicine tool that facilitates individualization of pharmacotherapy in accordance with genes influencing therapeutic response, side effects, and adverse events. It is an important application of personalized treatment that will help clinicians determine the right medication in the right dosage for their patients in a timely manner. Pharmacogenetic tests along with phenotypic information about patients and their underlying disease or condition have the potential to play an important role in drug therapy.

Pharmacogenetic testing or SNP profiling will help clinicians to tailor drug prescription and drug dosage to the individual and move away from the "one dose fits all" model (Marshall, 1997). Clinicians can pre-emptively opt to screen their patients before prescribing medication to maximize drug efficacy and minimize drug toxicity (Meyer, 2000; Roses, 2000; Wolf *et al.*, 2000; Issa, 2007). These tests have the potential to streamline the labour and time intensive drug discovery process, which includes crucial steps like drug development, drug testing and drug registration process. It will significantly help in reducing the time from the chemical synthesis of a drug, to introduction into clinical practice, and therefore reduce the cost of the entire drug development process (Roses, 2000). Knowledge obtained from the Human Genome Project coupled with all existing and new genomic technologies will provide scientists with a much broader range of specific targets with immense potential for therapeutic interventions (Issa, 2007). The wide array of existing and new genomic technologies like "genome wide association studies" (GWAS), "high-throughput technologies", "global gene expression analysis", "genome-wide functional analyses" and "gene expression monitoring" will help in the genetic analysis of numerous new molecular entities (NMEs) for drug efficacy and toxicity (Issa, 2002). This will enhance prospects for better and more effective therapies with minimal ADRs (Umamaheswaran *et al.*, 2014).

Taking all this into account, the pharmaceutical industry has over 100 clinical trials using pharmacogenetic data and pharmacogenetic strategies. Currently, there is about 417 active clinical trials that use extensive pharmacogenetic strategies (http://clinicaltrials.gov). The US FDA has a list of FDA-approved drugs and genes with pharmacogenomic information in their labelling (https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-

associations). The US FDA notes that pharmacogenetic tests can play an important role in avoiding ADRs by identifying responders and non-responders to medications and optimizing drug dosage. The Clinical Pharmacogenetics Implementation Consortium (CPIC) also has a list of drug–gene pairs, of which a few have US FDA pharmacogenetic label information where genetic testing is required, recommended, or considered 'actionable' (https://cpicpgx.org/genes-drugs/). In a study done by Sugarman, pharmacogenetic testing resulted in a 38% increase in the number of patients with changes to current medication. It also gave valuable genetic information that can be used to personalize future medical prescriptions for all patients (Sugarman *et al.*, 2016).

1.9 EXAMPLES OF PHARMACOGENETIC TESTS CURRENTLY USED IN CLINICAL SETTINGS

In high income countries, there has been a very significant increase in the use of pharmacogenetic tests before prescribing certain medicines. The US FDA has recommended the use of predictive genetic testing for specific drugs, but it is not mandatory. Pharmacogenetic tests are recommended because most drug treatments only have a response rate of 30% to 60% for the most common diseases (Spear *et al.*, 2001). Approximately 7% of hospital admissions are due to ADRs (Pirmohamed *et al.*, 2004). A significant proportion of treatment cessations is also because of ADRs. The toll this takes with respect to an individuals' suffering, high expenditure, and impact on their lives is a major burden of health care (Schwab *et al.*, 2012a).

When a health care provider considers prescribing a drug, knowledge of the patient's genotype will be very useful to identify a treatment plan, determine ideal drug dosage and assess the likelihood of benefit or toxicity. Pharmacogenetic tests, therefore, essentially help doctors identify potential responders and/or non-responders and/or fast/slow metabolisers before they prescribe a medicine. The following are a few examples of pharmacogenetic tests that will demonstrate their importance and relevance.

1.9.1 Glucose-6-phosphate dehydrogenase deficiency test

In some health settings, a pharmacogenetic test that has been accepted as a routine test is the glucose-6-phosphate dehydrogenase (G6PD) deficiency test. G6PD deficiency is an X-linked genetic disorder that affects about 400 million people worldwide (Anderle *et al.*, 2018). G6PD converts glucose-6-phosphate into glucose-6-phosphogluconolactone, the first step in the pentose phosphate pathway. G6PD plays an important role in protecting cells from oxidative stress/damage (Alharbi, 2015). Individuals with G6PD deficiency can develop severe jaundice in the neonatal period. Later on in their lives, if they are exposed to certain infections, drugs like the antimalarial primaquine, rasburicase (used in the treatment of hyperuricemia) and certain foods like fava beans, they can develop acute haemolytic anaemia (Bogari, 2016; Luzzatto *et al.*, 2016). In erythrocytes, G6PD enzyme regenerates nicotinamide adenine dinucleotide phosphate (NADPH), which subsequently produces glutathione. Glutathione protects red blood cells against oxidative attacks (Yang *et al.*, 2015).

X-linked disorders affect males and females differently. They are more likely to affect males than females because males only have a single copy of the gene. Since *G6PD* is an X-linked gene, hemizygous males are most likely to have severe forms of haemolytic crises (McDonagh *et al.*, 2012). If the X chromosome inherited from their mother is normal, men with X-linked disorders will pass on the faulty gene to all their daughters making them carriers of the disease. This cannot happen with sons because men only pass the Y chromosome to their sons.

Whether women with mutated *G6PD* will develop G6PD deficiency depends on a process called random X-chromosome inactivation (Beutler, Yeh and Fairbanks, 1962; Beutler, 1994). As women have two X chromosomes, certain disease traits on a X chromosome, may be masked by the normal gene on the other X chromosome, a process known as random X-chromosome inactivation (Alharbi *et al.*, 2013). Some women have favourable X-chromosome inactivation, where the affected X chromosome is silenced in most of the cells, thereby giving them sufficient G6PD enzymatic activity to avoid developing symptoms. Some females may have unfavourable X-chromosome inactivation, in which the unaffected X

chromosome is silenced in most of the cells, making them develop severe haemolysis and respond like men with G6PD deficiency (McDonagh *et al.*, 2012).

If drugs are to be prescribed in regions where G6PD deficiency is common, the WHO recommends testing of the patients in order to assess the risk of haemolysis in G6PD-deficient individuals ('Glucose-6-phosphate dehydrogenase deficiency. WHO Working Group', 1989). In order to address this and to minimise the occurrence of ADRs, the US FDA has warnings on the drug labels of drugs like primaquine, chloroquine, dapsone, rasburicase, avandaryl tablets, and glucovance tablets. Hence G6PD deficient individuals must be extremely cautious about what drugs they take. Other factors that contribute to drug-induced haemolytic anaemia in G6PD-deficient individuals are concomitant infections, high drug dosage, other genetic variants and other drugs taken in combination. This implies that drugs that are reported to cause haemolysis can be taken safely in individuals with G6PD deficiency provided they are administered with caution. It is advised that haemoglobin levels must be closely monitored especially when these drugs are administered in high doses and/or in combination with other drugs (Beutler, 1991; Youngster *et al.*, 2010).

1.9.2 Warfarin, CYP2C9 and VKORC1

Warfarin is an oral anticoagulant prescribed for the prevention and management of thromboembolic events that can culminate in stroke, deep vein thrombosis and pulmonary embolism (Daly *et al.*, 2003). Since warfarin has a narrow therapeutic range, identifying optimal warfarin dose at the beginning of treatment is quite challenging as there is no such ideal warfarin dose, because an appropriate dose in a patient may cause a haemorrhagic event in another. Approximately 1% of patients who have been administered warfarin die due to bleeding complications associated with warfarin. Minor bleeding complications are seen in 15% of patients treated with warfarin (Takahashi *et al.*, 2006). Warfarin efficacy is determined by how an individuals' anticoagulation can be maintained within an acceptable range without any risk of bleeding. It is primarily metabolised via oxidation in the liver by *CYP2C9*. Warfarin exerts its anticoagulant effect by inhibiting the protein vitamin K epoxide reductase complex, subunit 1 (VKORC1) (Cooper, Johnson, *et al.*, 2008; Takeuchi *et al.*, 2009; Biswas *et al.*, 2018).

Two SNPs in *CYP2C9* and one SNP in *VKORC1* play important roles in determining the effect of warfarin therapy on coagulation and in determining optimal dosing (Cooper, Johnson, *et al.*, 2008; Biswas *et al.*, 2018).

CYP2C9 has many alleles. *CYP2C9*1* is the wild type variant and *CYP2C9*2* (430C > T and rs1799853) and *CYP2C9*3* (1075A > C and rs1057910) reach polymorphic frequencies in several populations (Tabrizi *et al.*, 2002; Takahashi *et al.*, 2003; Cooper, Johnson, *et al.*, 2008; Takeuchi *et al.*, 2009; Biswas *et al.*, 2018). *CYP2C9*1* homozygotes or heterozygotes metabolise warfarin at standard doses. *CYP2C9*2* reduces warfarin metabolism by 30%, and *CYP2C9*3* reduces warfarin metabolism by 90%. *CYP2C9**2 and *CYP2C9**3 hence decreases warfarin maintenance dose requirement in patients and therefore patients homozygous or compound heterozygous for the **2* or **3* variants will metabolise warfarin less efficiently (Cooper, Johnson, *et al.*, 2008). In African-American populations *CYP2C9*2* and *CYP2C9*3* variant alleles have been reported at 3.3% and 2.3% (Dreisbach *et al.*, 2005). *CYP2C9*2* is not commonly reported in Asian populations. Whereas *CYP2C9*3* is found in frequencies of 1.1% to 6.8% (García-Martín *et al.*, 2006). A greater risk of bleeding is seen in patients with the alleles *CYP2C9*2* and *CYP2C9*3*. Hence a lower drug dose is required (Hillman *et al.*, 2005). Studies have shown that *CYP2C9* polymorphisms can be used to optimize warfarin treatment (Hillman *et al.*, 2005; Caraco, Blotnick and Muszkat, 2008).

In the promoter region of *VKORC1* at position, -1639, the common G allele is replaced by the A allele (g.-1639G > A; rs9923231) (Cooper, Johnson, *et al.*, 2008;Takeuchi *et al.*, 2009). This polymorphism lowers *VKORC1* gene expression and thereby lowers the production of the protein vitamin K epoxide reductase complex subunit 1 enzyme. This leads to warfarin sensitivity (Kudzi *et al.*, 2016). Hence people with the A allele produce less VKORC1 than those with G allele and therefore need lower warfarin doses to inhibit VKORC1 and produce an anticoagulant effect. *VKORC1* polymorphisms have been reported in 37% of Europeans and 14% of Africans (Rieder *et al.*, 2005). *CYP2C9*2*, *CYP2C9*3* and *VKORC1* promoter mutations together account for about 40–63% of variation in therapeutic warfarin dose (Rieder *et al.*,

2005; Sconce *et al.*, 2005). Therefore, pharmacogenetic testing is important to predict warfarin dosing along with other genetic and clinical factors.

Therefore, testing genotypes at the *CYP2C9* and *VKORC1* loci is highly recommended before prescribing warfarin as patients with a 'low dose' phenotype and genetically mediated sensitivity will require lower doses of warfarin to achieve therapeutic anticoagulation and avoid ADRs. The use of this genetic information is potentially useful in determining the correct start dose of warfarin in patients (Cooper, Johnson, *et al.*, 2008; Takeuchi *et al.*, 2009; Bader *et al.*, 2016; Biswas *et al.*, 2018). In African populations, because these alleles occur at lower frequencies, the test would have less of an impact on guiding therapy (Schelleman *et al.*, 2010; Kudzi *et al.*, 2016).

In addition to variants in *CYP2C9* and *VKORC1*, other factors like age, sex, body mass index, diet, concomitant drug therapy and ethnicity also affect warfarin dosage (Van Booven *et al.*, 2010). These factors can predict up to 60% of the variability in warfarin dosage in patients (Schwab *et al.*, 2011). Therefore both genetic and non-genetic factors have to be incorporated into warfarin treatment algorithms (Schwab *et al.*, 2011).

1.9.3 Imatinib Mesylate

Imatinib mesylate (Glivec, Gleevec or STI571) is an anti-cancer drug used in the treatment of chronic myeloid leukaemia (CML). It is a competitive tyrosine kinase inhibitor that has also been shown to be effective in the treatment of advanced gastrointestinal stromal tumours (GISTs)(Dulucq *et al.*, 2010). Neutropenia, superficial oedema, nausea, muscle cramps and rashes are few of the ADRs caused by IM (Baccarani *et al.*, 2009).

CML results from a rearrangement of DNA between chromosomes 9 and 22. The hybrid chromosome made is called the 'Philadelphia chromosome'. In this chromosome the ABL gene found on chromosome 9 is translocated to chromosome 22. This rearrangement causes the *ABL* and *BCR* genes to join together and make a fusion gene called *BCR-ABL* (Dulucq *et al.*,

2010). *ABL* produces an enzyme, tyrosine kinase, that under normal circumstances conveys signals between cells to control growth and division. *BCR-ABL* also encodes a tyrosine kinase, the enzymatic activity of which is not well regulated (Dulucq *et al.*, 2010) leading to large numbers of white blood cells being produced because they divide quickly and continuously. This, in turn, leads to characteristic symptoms of CML like fatigue, appetite loss and fever (Barratt *et al.*, 2017).

Glivec® can recognise and target this specific BCR-ABL fusion protein. Hence it can be used to block cancer cell growth in leukaemia patients who have the Philadelphia chromosome. The standardised drug dose range for Glivec has a fourfold inter-patient variability reported both in the systemic exposure for a given dose and in the dose required to achieve a specific drug activity level (Peng *et al.*, 2004). A starting dose of 400 mg per day is the standard Glivec dose recommendation for CML treatment (Barratt *et al.*, 2017). Up to 50% of CML patients treated with Glivec discontinue treatment due to poor efficacy or ADRs when using this one-dose-fits-all approach (Gotta *et al.*, 2014). Therefore, they might have to switch to other treatments which can be more expensive or/and can have significant toxicities. Varying Glivec dosage levels can result in extreme toxicity or suboptimal anticancer effect. Polymorphisms in Glivec's ADME genes and transporter genes affect drug bioavailability, response to drug and variability in inter-patient PKs. Glivec is mostly metabolised by *CYP3A4* and *CYP3A5* (Peng *et al.*, 2005). *CYP3A5*3* allele (rs776746) is defined by the substitution A6986G, which introduces a premature stop codon (Kuehl *et al.*, 2001). Individuals who are homozygous for this allele have reduced CYP3A5 levels and reduced metabolic capacity. *CYP2C8* genotype also significantly affects Glivec metabolism and exposure in patients with CML. *ABCG2* variant, 421C>A, is associated with improved treatment outcomes in Asian CML patients (Barratt *et al.*, 2017). Since Glivec treatment outcome correlates with above mentioned markers of ADME gene and drug transporter variability, personalised imatinib dosing is expected to significantly improve treatment outcomes compared to the very standardized "one-dose-fits-all" approach. Hence predictive testing is recommended before prescribing Glivec to identify potential responders and those who are likely to develop ADRs.

1.9.4 Trastuzumab

Trastuzumab (Herceptin®) is an anti-cancer agent used in the treatment of breast cancer. It is seen as the "poster child" for personalized medicine (Issa, 2007). Trastuzumab is a genetically engineered humanized monoclonal antibody (MAb) that was approved by the US FDA in 1998 (www.fda.gov/cder/approval/index). It was developed after the discovery that patients with breast cancer cells over-expressed the human epidermal growth factor receptor-2 protein (HER-2/neu). This protein is over expressed in 25-30% of breast cancer patients (Vogel *et al.*, 2002; Issa, 2007).

Trastuzumab was initially approved for the treatment of treatment-naïve metastatic breast cancer with over-expressed HER-2 protein in combination with paclitaxel, another chemotherapeutic agent. Trastuzumab by itself, is recommended for the treatment of breast cancer that has metastasised in patients with increased HER-2 expression who do not respond to other chemotherapeutics (Vogel *et al.*, 2002; Issa, 2007). A third of people of European origin with breast cancer have a mutation that causes increased HER-2 levels. Those with this mutation are hence likely to respond well to trastuzumab, with a reduction in tumour size, slowing down disease progression and increasing survival (Kudzi *et al.*, 2011). Thus, the use of this genetic information is critical in determining the correct mode of treatment for HER-2 positive breast cancer as such patients will respond better to Trastuzumab.

1.9.5 Azathioprine and TPMT

Azathioprine (AZA) is an immunosuppressant which interferes with DNA synthesis and inhibits the proliferation of rapidly growing cells. It is used in patients undergoing organ transplantation to prevent rejection. It is used in the treatment of inflammatory and auto-immune disease, and certain types of cancers like leukaemia (acute lymphocytic leukaemia, ALL) and lymphomas. 6-mercaptopurine (6-MP) is a metabolite of AZA. Thiopurine methyltransferase (TPMT) is responsible for the metabolism of AZA and 6-MP (Gearry and Barclay, 2005; DiPiero, Teng and Hicks, 2015).

The *TPMT* gene has genetic polymorphisms that lead to almost 50-fold variation in enzyme activity between individuals (Gearry and Barclay, 2005). Twenty-eight polymorphisms have been identified in *TPMT*, most of which are associated with decreased enzymatic activity. Compared with people with normal TPMT activity, people with intermediate or low enzyme activity metabolise the drug very slowly (DiPiero *et al.*, 2015). Patients with TPMT deficiency are at an increased risk for drug-induced toxicity because of the accumulation of the cytotoxic 6-thioguanine nucleotide analogues, which cause severe to fatal haematological toxicity and myelotoxicity (Baker, 2003). Patients with high TPMT activity have increased drug clearance and hence may not benefit from the drug at the standard drug dose.

When TPMT-deficient patients are administered thiopurine drugs at standard doses, they will develop serious myelotoxicity (Kaskas *et al.*, 2003). When patients with intermediate TPMT activity are prescribed high doses of thiopurine drugs, they might be at increased risk of myelotoxicity (Seidman, 2003). Therefore, routine *TPMT* genotyping is recommended before prescribing thiopurine drugs like AZA and 6-MP, in order to decrease the risk of severe and possibly preventable ADRs and to identify patients who might benefit from higher doses (Baker, 2003).

1.9.6 Other pharmacogenetic tests

Some other pharmacogenetic tests that have been approved by the US FDA include cetuximab/panitumumab and *KRAS*; vemurafenib and *BRAF*; abacavir and *HLA-B*5701*; carbamazepin and *HLA-B*1502*; and thiopurines and *TPMT* (Meyer *et al.*, 2013).

The US FDA has published a table that has information about pharmacogenetic associations that are related to drug metabolising enzyme gene variants, drug transporter gene variants, and gene variants that are related to a predisposition for certain adverse events (https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations). Common pharmacogenomic based personalised medicine applications are listed in Table 1.7.

Table 1.7: Personalised medicine application based on pharmacogenomics (Issa, 2007)

Drug	**Polymorphic gene**
Phenytoin	*CYP2C9*
Tolbutamide, glipizide	
Flouxetine	*CYP2D6*
Codeine	
Thioridazine	
Strattera (atomoxetine)	
5-Fluorouracil	*DPYD*
Antiepileptic drugs	*MDR1* (*ABCB1*)
6-Mercaptopurine	*TPMT*
6-thioguanine	
Azathioprine	
β-agonists (e.g. albuterol)	β2–adrenergic receptors
β-blockers (e.g. propranolol)	*ACE*
Losartan	*AT-1*
Irinotecan	*UGT1A1*
Abacavir	*HLA*
Sulfonamides	*NAT*s
Isoniazid, Procainamide	
Hydralazine	
Imatinib mesylate	*BCR*/*ABL*
Trastuzumab	Her-2

1.10 EXPRESSION QUANTITATIVE TRAIT LOCI (eQTLs)

As discussed previously, genetic polymorphisms explain some of the variability in pharmacogene expression, response to drug therapy, likelihood of getting ADRs and clinical outcome of disease. These genetic polymorphisms impact the pharmacokinetics and pharmacodynamics of drug therapy, which are of ultimate importance to pharmacogenetics (Huang *et al.*, 2008).

Functional genetic variants primarily alter the function of the enzyme or other protein by changing its structure. Copy number variants alter the number of genes present, influencing the amount of enzyme/protein produced. Another class of variants called eQTLs affect gene expression and alter the amount of enzyme or other proteins involved in drug metabolism and transport, in a tissue specific manner (Stranger *et al.*, 2012). eQTLs are regions of the genome containing genetic variants (SNPs) that influence or are associated with gene expression variation (Glubb *et al.*, 2012; Nica *et al.*, 2013). eQTLs are also known as expression single nucleotide polymorphisms (eSNPs) (Franke *et al.*, 2009). This thesis is focused on the detection of eQTLs for ADME genes.

With genotyping becoming more affordable, it has become possible to analyse large groups or populations of unrelated individuals. This has resulted in the association of many genes and genetic polymorphisms with complex diseases (Manolio *et al.*, 2009; Milne *et al.*, 2017; Visscher *et al.*, 2017; Zhao *et al.*, 2017; Buniello *et al.*, 2019). However, for a significant proportion of these genes, it is not yet clearly understood what the biological effects of the variants are. Some of the variants could affect gene expression levels in a tissue-specific manner, and consequently the amount of protein produced by a cell. Further studies could help uncover the effects of the variants on the expression of specific genes, proteins, metabolites, and phenotypes.

With the advent of high-throughput and genome-wide measurement of gene expression in large populations it has become possible to link variation in gene expression to SNPs. It is essential to have both genotype and transcriptome data from the same individual to identify genetic variants associated with variation in gene expression (eQTLs).

eQTLs that exert regulatory effects on the expression of nearby genes are referred to as *cis* eQTLs (Stranger *et al.*, 2012). By definition, a *cis* or local eQTL is within 1 Mb on either side of the transcription start site (TSS) of the gene whose expression it modulates (Price *et al.*, 2008; Nica *et al.*, 2013; Wright *et al.*, 2014; Aguet *et al.*, 2017). *Trans* eQTLs exert regulatory effects

on the expression of genes at longer genomic distances (Stranger *et al.*, 2012). By definition, a *trans* or distant eQTL is at least 5 Mb downstream or upstream of the TSS of a gene whose expression it modulates (Price *et al.*, 2008; Nica *et al.*, 2013; Wright *et al.*, 2014; Aguet *et al.*, 2017). It is found that regulatory control mostly occurs locally in the vicinity of the genes (Göring *et al.*, 2007; Veyrieras *et al.*, 2008; Stranger *et al.*, 2012). Consequently, *cis*-eQTLs have been detected for many genes. Most strong eQTLs are found near target genes (Göring *et al.*, 2007; Veyrieras *et al.*, 2008; Stranger *et al.*, 2012).

Large sample datasets increase the power of eQTL studies and hence with increased power, the number of genes with identified eQTLs will also increase (Veyrieras *et al.*, 2008; Nica *et al.*, 2013). eQTL studies show that variation in gene expression levels is widespread, highly heritable and can be genetically mapped (Gilad *et al.*, 2008). These attributes help in the identification of eQTLs (Gilad *et al.*, 2008).

Most genetic variants in genes associated with human diseases do not directly cause the disease or confer protection from the disease. Most of the novel highly replicated disease trait associations do not have supporting experimental data to explain what role the given susceptibility gene plays in the onset and/or progression of disease (Schadt *et al.*, 2008). Even when the susceptibility genes are well studied, like the Alzheimer disease susceptibility gene *APOE*, understanding how this gene endows disease susceptibility can take years (Peacock *et al.*, 1993). Some of the associated variants act on intermediate phenotypes like gene expression, protein expression, protein state and metabolite levels. They alter gene expression by their effect on RNA transcription or splicing (O'Brien *et al.*, 2018). These intermediate phenotypes then go on to induce changes in higher-order disease associated traits, ultimately culminating in disease. Intermediate molecular phenotypes can differ as a consequence of genetic variation and may associate with changes in disease traits. Most SNPs detected that are associated with disease traits from GWAS do not appear to affect protein sequence (Schadt *et al.*, 2008; Sebastiani *et al.*, 2009; Nicolae *et al.*, 2010). Some of these SNPs are thought to regulate gene activity either at the transcript level or link to other genetic polymorphisms involved in regulatory roles (Schadt *et al.*, 2008). Therefore, by identifying the

different intermediate phenotypes, the genes that are affected by eQTLs can be identified and further validated. They will help understand the molecular networks in which these genes operate and how changes in these networks will lead to changes in disease traits (Schadt *et al.*, 2008).

eQTL mapping is a reliable approach to identify common genetic variants that affect gene expression levels (Deelen *et al.*, 2015). It can result in the same regulatory region and variant being linked to one gene in one tissue and to a different gene in another tissue (Nica *et al.*, 2013). By combining genome-wide genotyping with studies on gene expression (transcriptome profiling), variants associated with altered gene expression can be mapped on a genomic scale as eQTLs (O'Brien *et al.*, 2018).

eQTL mapping can help give us a clearer understanding of the underlying mechanism of disease. It has been successfully applied to study diseases like celiac disease and asthma (Moffatt *et al.*, 2007; Hunt *et al.*, 2008). Genetic variants associated with these diseases affect gene expression either in *cis* or in *trans*. Such genetic variant-gene expression associations could provide insight into the biological pathways affected by these diseases. One of the major advantages of eQTL mapping is that it facilitates the identification of new functional loci without needing any previous knowledge of regulatory elements (Nica *et al.*, 2013). One of the most effective ways to discover gene regulation networks is by identifying eQTLs (Brem *et al.*, 2005). Identifying eQTLs will measure variance in RNA abundance and map the genetic loci affecting the expression of mRNA (Jansen *et al.*, 2001).

1.11 THE GENOTYPE-TISSUE EXPRESSION PROJECT

The Genotype-Tissue Expression project (GTEx) aims to build a comprehensive public resource focussing on the study of tissue specific gene expression and regulation. The GTEx project established a freely available database to study the relationship between genetic variation and gene expression regulation (Lonsdale *et al.*, 2013; Aguet *et al.*, 2017). It provides the most comprehensive survey of tissue-specific gene expression which greatly enhances the

ability of researchers to study the impact of genetic variation on gene expression levels (Consortium, 2013).

Though significant associations have been established between common genetic variants and different human traits from GWAS, the functional effects and underlying mechanisms of majority of these variants are still unexplained (Altshuler *et al.*, 2008; Hindorff *et al.*, 2009; Lonsdale *et al.*, 2013). Complex diseases, like diabetes, are often caused by dysfunction in multiple cell types like pancreas, adipose, and skeletal muscle thereby making it difficult to narrow what the causal tissue(s) are for any disease or GWAS trait associated with it (Voight *et al.*, 2010). Thus, in order to comprehend the functional association of GWAS loci with disease aetiology, it is pivotal to understand what role the regulatory variants have in the relevant tissues. GTEx addresses this limitation by establishing a sample and data resource to study the relationship between genetic variation, tissue-specific gene expression and disease aetiology (Consortium, 2013).

The stated goals of the GTEx project are the following (Consortium, 2013):

- To create a data resource to enable the systematic study of both genetic variation and gene expression regulation in various reference human tissues.
- To provide the scientific community with a biospecimen resource that includes tissue, nucleic acid, and cell lines to conduct further studies required to determine other molecular phenotypes.
- To support and disseminate the results of a study of the ethical, legal, and social issues related to donor recruitment and consent.
- To support the development of novel statistical methods for the analysis of human eQTLs by themselves and in the context of other molecular phenotypes.
- To make the data publicly available to the scientific community as expeditiously as possible.

- To support the dissemination of knowledge, standards, and protocols related to the above five goals that were developed during the project.

Besides information on *cis*- and *trans*- eQTLs, information on allele specific expression analysis and splicing eQTLs can also be obtained from the GTEx portal (https://gtexportal.org/) (Consortium, 2013, 2015).

1.12 THE ROLE OF eQTLs IN PHARMACOGENOMICS

eQTLs and their direct association with gene expression can provide potential new links between the genotype and phenotype that can be clinically or functionally validated. These new links will also help explain the importance of a specific gene of interest. This information will enhance personalised medicine by providing vital information that can lead to the identification of new disease biomarkers, new therapies, and diagnostic procedures. These biomarkers can then be put to good use in the detection and treatment of diseases (Huang *et al.*, 2008).

Genetic markers that are associated with both eQTLs and disease status are important because if one allele is more frequent in cases than in controls, and at the same time affects the expression of a nearby gene, the genetic marker's causality can be established (Nica *et al.*, 2013). When GWAS results are integrated with gene expression data, genes, and pathways whose disruption causes disease can be discovered. This is however only possible when the tissue of expression is relevant to the interrogated disease (Nica *et al.*, 2008).

Incorporating eQTL analyses with GWAS results can help propose new disease candidate genes. A series of strongly correlated SNPs in chromosome 17q23 was found to be associated with childhood asthma (Moffatt *et al.*, 2007). This 200 kb region has 19 genes, none of which had a clear role in childhood asthma. *ORMDL3* (ORM1-like 3) was found to be the best candidate for further functional studies because when expression studies were performed on

lymphoblastoid cell lines derived from the same families, it showed that the most relevant GWAS SNPs explained about 29.5% of the variance in transcript levels of *ORMDL3*.

Expression data has also helped in interpreting some of the association signals for Crohn's disease. The findings of a GWAS study led to the mapping of a 1.25 Mb gene desert region on chromosome 5 that has multiple susceptibility loci (Barrett *et al.*, 2008). eQTL studies showed that a few of these loci act as *cis* regulators of *PTGER4* (prostaglandin E receptor 4), a gene that is 270 kb away from the associated region (Libioulle *et al.*, 2007). Several other analyses which support the use of eQTL data in interpreting GWAS results have been done for height, systemic lupus erythematosus, type I diabetes and bipolar disorder (Consortium, 2007; Hakonarson *et al.*, 2007; Gudbjartsson *et al.*, 2008; Hom *et al.*, 2008; Nica *et al.*, 2013).

Though eQTLs can be used as biomarkers for both disease association and in predicting clinical outcomes, this study will only focus on the potential role of eQTLs in pharmacogenomics. In the context of clinical pharmacogenomics, the liver, a metabolically active tissue that is critical to many core physiological processes, is the most relevant tissue in which to perform eQTL analyses. Liver is crucial for many human metabolic processes (Wang *et al.*, 2014). It plays vital roles in metabolic processes like maintaining homeostasis and health, synthesis of essential serum proteins, production of bile and its carriers, and the regulation of nutrients (Innocenti *et al.*, 2011). It is the most important organ for drug metabolism and drug elimination as it is involved in the metabolism of almost 75% of the 200 most extensively prescribed drugs (Wienkers *et al.*, 2005).

Many ADME genes are strongly expressed in the liver. It has therefore been postulated that it is unlikely that eQTLs of ADME genes will be detected in any other tissues (Glubb *et al.*, 2012). Altered xenobiotic metabolism by genetic variants in ADME genes or by eQTLs that affect ADME gene expression will affect the systemic availability of the drugs, thereby impacting their pharmacologic effects (Lin, 2007). Liver eQTLs will also help understand associations of SNPs with clinical phenotypes. This provides us with a sound rationale that

eQTLs for relevant genes in the liver can be used for selection of candidate SNPs for further pharmacogenomic research. Liver eQTL data has the potential to help us understand how SNPs are significantly associated with clinical traits (Glubb *et al.*, 2012). eQTLs identified for any of the ADME genes that are expressed in the liver can be used to test pharmacogenetic associations with clinically relevant phenotypes of the multiple drugs that are metabolised by the ADME gene in question. (Glubb *et al.*, 2012). Information about all known drugs that are metabolised by the ADME genes is available on "The Pharmacogenomics Knowledge Base" (PharmGKB; www.pharmgkb.org). PharmGKB "curates pharmacogenetic findings from the literature" and has information on "associations between germline genetic variants and drug-related phenotypes."

Majority of drug metabolism occurs in the liver (Innocenti *et al.*, 2011). Therefore, it was decided that it would be optimal to study eQTLs of ADME genes in the liver. It must be noted that gene expression signatures are cell-type specific (Savci-Heijink *et al.*, 2019; Henry *et al.*, 2021), but at the time of this study, due to the unavailability of datasets on hepatocytes in the public domain, it was decided that it would be sufficient to study eQTLs in the liver. But since the liver is a collection of different cells types like hepatocytes, hepatic stellate cells, Kupffer cells, and liver sinusoidal endothelial cells, to name a few (Ding *et al.*, 2016), we are making an implicit assumption that studying the expression levels in the liver is sufficient. Moreover, mapping of eQTLs in liver has demonstrated its power for understanding inter-individual variability in disease aetiology and response to therapeutic treatments (Schadt *et al.*, 2008). Liver eQTL research can therefore be a stepping stone for pharmacogenetic research (Glubb *et al.*, 2012). Understanding the basis of genetic variation in hepatic gene expression will significantly improve our understanding of pharmacogenomics.

Since DNA genotyping is clinically more practical than gene expression measurements in the liver and because eQTLs directly link gene variants with gene expression, identifying eQTLs could lead to the development of important biomarkers for drug efficacy and ADRs. Hence the identification of genetic determinants for genes whose expression play a significant role

in drug response/toxicity can be clinically very useful as pharmacogenetic markers (Huang *et al.*, 2008).

An eQTL analysis done by Hernandez *et al.* (2017) identified rs4889606, a strong *cis*-eQTL for *VKORC1,* near the 3' end of the gene. The combination of genetic factors like *VKORC1* -1639G>A (rs9923231) and *CYP2C9* (**2* and **3*) alleles, and non-genetic factors like age, sex, weight and concomitant drug use that are found to influence warfarin metabolism explain around 55% of the variation in the required maintenance dose for Caucasians, but only about 25% for African–Americans (Wen *et al.*, 2008; Limdi *et al.*, 2010). The predictive power of several warfarin dosing algorithms, comprising of genetic and non-genetic covariates from studies accounts for no more than 60% of the variance in warfarin dose requirements in ambulatory patients (Wen *et al.*, 2008)**.** Inclusion of the genotype for the eQTL, rs4889606, along with previously identified genetic factors (the *CYP2C9* alleles, rs9923231 genotype) and other clinical variables, explained about 31% of inter-patient variability in warfarin dose requirement in African-Americans (Hernandez *et al.*, 2017). Thus, the inclusion of the eQTL improved our understanding of variability in warfarin dose requirement in African-Americans.

Acenocoumarol is a vitamin K antagonist oral anti-coagulant that is recommended to prevent cardiovascular events. It is a racemic mixture, and the R-enantiomer, R-acenocoumarol, is metabolised by *CYP1A2*, *CYP3A4*, *CYP2C9* and CYP2C19. The S-enantiomer, S-acenocoumarol, is metabolised by *CYP2C9.* It works by inhibiting vitamin K epoxide reductase (VKOR) (Ufer, 2005). The main genes involved in the pharmacogenetics of a vitamin K antagonist oral anti-coagulant are: *VKORC1, CYP2C19* and *CYP4F2*. Variants in these genes are included in pharmacogenetic algorithms to predict the drug dose requirement in each patient, depending on their genotype and clinical variables. Acenocoumarol has a very narrow therapeutic profile with high variability in dose requirements. Genetic and environmental factors also contribute to variability in drug dose (Bodin *et al.*, 2005). An eQTL, rs10871454 was found in the Syntaxin-4 (*STX4*) gene and was significantly associated with acenocoumarol dose variance in different populations. This eQTL acts by decreasing the hepatic expression of VKORC1 mRNA, thereby decreasing the amount of the drug. This association has been

replicated and when incorporated in the pharmacogenetic algorithm to predict drug dose, our understanding of drug dose variability improves (Cullell *et al.*, 2018).

Hence, it is postulated that eQTLs of ADME genes can explain some of the variance in drug dose requirements that contribute to the currently unexplainable portion of variability in drug response. Treatment response predictions/algorithms could be enhanced by adding eQTLs in combination with existing knowledge of genetic variations and other clinical factors. eQTLs of ADME genes expressed in the liver may help explain the associations of SNPs in these ADME genes with optimal drug dosage.

If eQTLs are found for any of the ADME genes, they can be used to test relevant pharmacogenetic associations with any of the many drugs that are metabolised by the corresponding enzyme(s). Therefore, the aim of this study is to identify eQTLs that potentially affect the expression of ADME genes in the liver and thereby improve our understanding of the role of these genetic effects on drug therapy.

1.13 PREVIOUSLY PUBLISHED eQTL STUDIES OF IMPORTANCE

This section focuses on two studies that are relevant to this thesis. The first one is a study done on liver eQTLs in Caucasians by Schröder (Schröder *et al.*, 2013). The pipeline was built using datasets from this study. The second one is a study on liver eQTLs in African-Americans (AAs) by Zhong (Zhong *et al.*, 2020).

In the first study, 149 surgically removed livers from Caucasian donors were investigated to identify associations between genetic polymorphisms and gene expression. Special emphasis was given for genes involved in the absorption, distribution, metabolism, and excretion of drugs. Pre-processing and quality control done on the transcriptome and genotype dataset is discussed in detailed in Section 2.4 as this dataset was used for building the pipeline. GenABEL, a R package, was used to identify *cis*- and *trans*-eQTLs by testing for all possible combinations of SNPs and expression traits. An additive linear model was used to identify

eQTLs. Sex, age, smoking status, consumption of alcohol, C-reactive protein level, presurgical medication, cholestatic liver disease and diagnosis were used as covariates for this analysis.

This study identified 1226 significant eQTL associations of which 1179 were *cis*-eQTLs and 47 were *trans*-eQTLs between 1163 SNPs and 371 different expression traits. Since the focus was on ADME genes, the identified eQTLs were filtered according to their relevance in ADME processes relevant to drug disposition and toxicity. Subsequently, either the SNP or the gene or both gene and SNP had to be relevant to an ADME gene. This resulted in the identification of 95 significant eQTLs of which 89 were *cis*-eQTLs and 6 were *trans*-eQTLs. Of the 95 eQTLs that were identified, 21 associations were validated by comparison to a previously published study and 74 ADME associations were uniquely identified by Schröder *et al*. Of the 21 associations that were validated by comparison to a previously published study three were exactly replicated eQTLs and 18 were matches by LD. Of the 24 ADME associations uniquely identified by Schröder *et al*, 68 were *cis*-eQTLs and 6 were *trans*-eQTLs. The low reproducibility between studies was attributed to differences in technology platforms used for genotyping and gene expression profiling, origin of the sample, number of samples analysed and available medical information on the sample donors.

In the second study, 68 AA hepatocyte cultures were used to identify AA specific hepatocyte eQTLs. Hepatocyte cells were either bought from commercial companies or isolated from cadaveric livers. Livers with active cancer or a history of hepatocarcinoma were excluded. Very briefly, PLINK was used to identify individuals with discordant sex information and duplicated or related individuals. Samples that did not cluster along with the AA samples on the PCA were excluded from the dataset. This resulted in the exclusion of 6 samples, leaving 62 samples in the dataset. SNPs on the sex and mitochondrial chromosome, SNPs that might introduce flip-strand issues, SNPs with missingness greater than five percent and SNPs that deviate from the Hardy-Weinberg equilibrium were removed from the genotype data leaving 674,996 SNPs. Genotype data was phased using SHAPEIT and IMPUTE2 was used for imputation using all reference populations from 1000 Genomes phase 3. Post imputation,

SNPs were excluded for minor allele frequency < 0.05, imputation quality scores < 0.8, and HWE p-value < .00001, leaving 7,180,502 SNPs in the analysis.

Only samples with RNA integrity number (RIN) score of greater than 8 were considered for RNA sequencing. FastQC (v0.11.2) was used to assess raw reads from FASTQ files. Only FASTQ files with per base sequence quality threshold of > 20 across all bases were used. STAR 2.5 was used to align reads to human Genome sequence GRCh38 and Comprehensive gene annotation (GENCODE version 25). SAMTools 1.2 was used to index uniquely mapped reads only. read_NVC.py, read_GC.py and geneBody_coverage.py from RNA-SeQC were used to assess nucleotide composition bias, GC content distribution and coverage skewness of the mapped reads respectively. Samples with normally distributed GC content, and, without any nucleotide composition bias and coverage skewness were reserved. The distribution of bases within the transcripts was evaluated by Picard CollectRnaSeqMetrics. Finally, nucleotide distribution within specific regions of the genome was measured and only samples with more than 80% of bases aligned to exons and UTRs were included in the dataset for analysis. RNA-SeQC was used to evaluate gene expression after the reads were mapped to genes that were referenced with gene annotation from GENCODE v25.

Reads were mapped to genes referenced with Comprehensive gene annotation (GENCODE version 25) to evaluate gene-level expression using RNA-SeQC. Bioconductor package DESeq2 (version1.20.0) was used to supply HTSeq raw counts needed for gene expression analysis. Log transformation was used to normalize counts. DESeq2 was used to perform PCA. Sample expression pattern was visualized with PC1 and PC2 and two samples with distinct expression patterns were excluded from the dataset.

edgeR was used to normalize gene expression by trimmed mean of M-values normalization method (TMM). Transcript per million (TPM) was calculated by first normalizing the counts by gene length and then by read depth. Gene expression values were filtered based on expression thresholds < 0.1 TPM and ≤ 6 reads in a minimum of 20% of the samples. Inverse

normal transformation was used to normalize each gene's gene expression value across samples. Gene coordinates were then remapped to hg19/GRCh 37 (GENCODE version 19).

Probabilistic Estimation of Expression Residuals (PEER) factors were calculated to correct for both hidden and measured confounders of gene expression analysis and to normalize gene and exon expression values using PEER R package. Matrix eQTL was used to identify eQTLs using the following covariates: top 10 PEER factors, sequencing platform, batch, first PC calculated from pre-imputation LD-pruned genotypes and sex.

Local ancestry of the AA genotypes was estimated with RFMix using YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah Residents with Northern and Western European Ancestry from the CEPH collection) samples from 1000 Genomes phase 3 as the reference populations. LAMatrix was used for mapping eQTL with local ancestry information. An eQTL was identified as significant if the false discovery rate (FDR) is less than 0.05.

This study identified 277 eGenes and 19,770 significant eQTLs in AAs. Conditional analysis identified 240 secondary eQTLs for *HCG4P7* and *PPIL3* of which 217 were not identified in the primary analysis. 137 genes whose expression was associated with local ancestry at FDR <0.1 and 65 genes whose expression was associated with local ancestry at FDR <0.0.05 was found. This supported the use eQTL mapping with local ancestry adjustments. The expression of *GSTA2*, which has a potential role in acute anthracycline-induced cardiotoxicity for which AAs are at a heightened risk, was found to be associated with local ancestry. By using LAMatrix an additional 1179 eQTLs were identified as compared to eQTL mapping by MatrixEQTL. eQTLs identified by local ancestry adjustment showed a greater enrichment in histone active markers like H3K27ac and H3K4me1 in the Roadmap liver tissue.

When AA hepatocyte eQTLs were compared to eQTLs found in the GTEx liver dataset(v7) that were filtered at by MAF>0.05, of the 277 eGenes identified in AAs, 210 were also eGenes in the GTEx liver dataset. Of the 210 eGenes that overlapped between both GTEx and this study,

only 25 were found to share the same lead eQTL (the most significant eQTL for each gene) implying that the genetic architecture of gene expression in European samples (that was predominantly used in GTEx) is different from that in AA samples. When effect sizes of the overlapping eQTLs between this study and GTEx were compared, a strong correlation was found that indicates that genetic regulation is shared between Caucasians and AAs. The effect sizes of the overlapping eQTLs were found to be higher in AAs hepatocytes compared to the liver dataset used in GTEx. AA-specific eQTLs were found to have higher allele frequencies in African population (AFR) of the 1000 Genome Project compared to European population (EUR). Fst was used to measure the difference in allele frequencies between populations and it was found that AA specific eQTLS have higher Fst values compared to that of overlapping eQTLs. This highlights the importance difference in allele frequencies has to population specific identification of eQTLs.

AA-specific eQTLs were found to have larger Fst compared to overlapping eQTLs which demonstrates the contribution of allele frequency differences to population-specific eQTL discoveries. The intersection of SNPs associated with traits and eQTLs identified in this study was determined by comparing GWAS variants or their tagging SNPs from the GWAS catalog with the identified eQTLs. 721 GWAS associations were found to intersect with AA eQTLs, which provided evidence that eQTL genes play roles in different traits. The number of AA eQTLs that intersected with GWAS associations was found to be significantly less per loci thereby suggesting that when GWAS associations intersect with AA eQTLs, the number of potential causal variants at that site can be narrowed down.

exon QTLs were mapped in this study to uncover the genetic regulation of alternative exon usage. 1,284 exon segments with significant genetic regulation (eExons) and 75,002 exon QTLs were identified. Among the 730 genes identified for all eExons, 223 genes were also identified as eGenes in eQTL mapping. 507 genes of the 730 were found to be unique to the exon QTL analysis. 51.2% of the exon QTLs (51.25) were also identified as eQTLs.

This study was the first eQTL study conducted in liver in AAs, an underrepresented population (Innocenti *et al.*, 2011). The eQTLs in this study were discovered despite the small dataset because of the much higher allele frequency of the eQTLs in populations of African ancestry. This study highlights the importance of mapping eQTLs in populations of African ancestry and other diverse populations with greater genetic diversity. Differences in allele frequency and LD structure can help in the fine-mapping of novel regulatory variants.

1.14 THE IMPORTANCE OF PHARMACOGENOMIC RESEARCH IN AFRICA

Pharmacogenomics research in Sub-Saharan African populations is vital, as there is a high prevalence of communicable diseases like HIV/AIDS, tuberculosis, to cite a few and non-communicable diseases such as hypertension, diabetes and stroke (Kudzi *et al.*, 2016). HIV/AIDS is a major cause of death in young adults, tuberculosis affects all age groups, and over 10% of children in sub-Saharan Africa die before the age of 5 due to infectious diseases (malaria, respiratory infections and diarrhoea) compared to <1% in high-income countries (Teo *et al.*, 2010). The prevalence of non-communicable diseases like hypertension, diabetes, and obesity in people of African ancestry is higher than in people of European ancestry.

African populations are sub-divided based on both geographic location and language (Rotimi *et al.*, 2016). A genetic study based on available data on Africans at the time identified eleven distinct ancestries in Africa, thereby showing high genetic substructure (Tishkoff *et al.*, 2009; Shriner *et al.*, 2014; Gurdasani *et al.*, 2015), although this conclusion will likely change as more data becomes available on additional African populations (Choudhury *et al.*, 2020). Genetic differentiation, between the different African ancestries can be higher than genetic differentiation measured between non-African ancestries (Shriner *et al.*, 2014). African populations are known to harbour the greatest genetic diversity (Campbell *et al.*, 2010; Choudhury *et al.*, 2018; Thami *et al.*, 2019). They are characterized by extensive population substructure and less LD among loci compared to non-African populations. African populations have shorter blocks of LD because they have had more time for recombination to decrease levels of LD being the oldest population in the world (Choudhury *et al.*, 2018; Thami *et al.*, 2019). The complex population history of the continent, the dramatic variation

in climate, diet, and exposure to infectious disease, result in higher than normal levels of genetic and phenotypic variation in African populations (Campbell *et al.*, 2008). Shorter LD blocks in Africans makes characterizing patterns of LD quite challenging. The high genetic diversity also attributes to African and African-American populations having more recombination hotspots than non-Africans.

Extensive population substructure seen in Africa results in different study sites in Africa having different allele frequencies and patterns of LD. This drastically reduces the likelihood of reproducing associations in multi-centre studies. However, it probably will be relatively easy to localize the causal variant in Africans because of low levels of LD with neighbouring SNPs (Teo *et al.*, 2010). The above-mentioned reasons show how important it is to map and understand the genetic make-up of Africans so that maximum drug efficacy and minimum ADRs is achieved. Thus, finding relevant eQTLs in populations from sub-Saharan African will be beneficial in understanding the role of variation and the unique pharmacogenomic make-up of individuals.

1.15 RATIONALE AND MOTIVATION FOR THE STUDY

It is well-known that genetic polymorphisms and environmental factors influence how effective a drug is and the individuals' propensity to develop ADRs (Wilke *et al.*, 2007). However, 90% of variants that are found to be significantly associated with complex phenotypes are found in non-coding regions of the genome (Giral *et al.*, 2018). Regardless of advances in understanding drug response variation, the major causes of this variation remain unexplained. Extensive in-depth knowledge of genetic variations, pre-existing clinical factors and environmental variables currently only explains a small proportion of variance in drug response. There still is a large gap in our knowledge of what exactly causes inter-individual differences in drug response.

Change in gene expression levels has been found to be a major mechanism that underlies drug response variation and susceptibility to disease (Roden *et al.*, 2011). eQTLs modulate gene expression levels and in this study, I aim to find eQTLs for ADME genes which could

improve our understanding of drug-dose variance that contributes to the currently unexplainable portion of variability in drug response and propensity to ADRs. Treatment response predictions/algorithms can also be improved significantly by adding eQTLs in combination with existing knowledge of genetic variations and other clinical factors. For example, an eQTL identified near the 3' end of the *VKORC1*, a gene is associated with warfarin dosing, could explain some of the variance in drug dose requirements that contributes to the currently unexplainable portion of variability in drug response (Schadt *et al.*, 2008). This unexplainable portion of variability in drug response could not be explained by a combination of genetic (markers in *VKORC1* and *CYP2C9* gene: *VKORC1* -1639G>A, *CYP2C9 *2* and **3*), and non-genetic factors (e.g., age, sex, weight and amiodarone use) (Limdi *et al.*, 2010). The predictive power of a number of warfarin dosing algorithms, comprising genetic and non-genetic covariates from studies accounts for no more than 60% of the variance in warfarin dose requirements in ambulatory patients (Gardiner *et al.*, 2006; Limdi *et al.*, 2008; Biswas *et al.*, 2018).

If eQTLs are found for any of the ADME genes, they can be used to test relevant pharmacogenetic associations with any of the many drugs that are metabolised by the corresponding enzyme(s). Therefore, the aim is to identify eQTLs that affect the expression of ADME genes with special reference to the liver and to improve our understanding of pharmacogenetics of drug therapy.

1.16 STUDY AIMS AND OBJECTIVES

The aim of this study is to improve our understanding of the effect eQTLs have on drug therapy outcomes. This will be achieved by studying the effect eQTLs have on ADME genes. Since the effect of eQTLs are tissue specific and the aim is to study the effect eQTLs have on ADME genes, the focus is on liver tissue as it is the most important organ for drug metabolism. The main objectives for this research are:

Objective 1: To develop a robust pipeline to identify eQTLs from datasets that include genome and transcriptome data on the same set of participants.

Objective 2: To identify *cis*- and *trans*-eQTLs in liver tissue for ADME genes.

Objective 3: To perform in-silico analyses of the predicted function of the eQTLs.

CHAPTER 2 METHODS AND DATA ANALYSIS

eQTL analysis is a computationally intensive task that involves testing for association of many SNP-transcript pairs. Genotype data and corresponding gene expression data are a prerequisite for eQTL analysis. Both genotype and gene expression data needed for this study were freely available in the public domain and were downloaded from the Gene Expression Omnibus (GEO) database from the National Center for Biotechnology Information (NCBI) resource (https://www.ncbi.nlm.nih.gov/geo/). GEO is a public functional genomics data repository that stores array and sequence data. It has tools that help users query and download experimental data and curated gene expression profiles.

2.1 PUBLIC DATA ACCESSED

The only suitable publicly available dataset was from a study by Schröder *et al.*, 2013. This study collected liver tissues and corresponding blood samples from 150 patients of European ancestry who underwent liver surgery at Campus Virchow (University Medical Center Charité, Humboldt University, Berlin, Germany). Of the 150 patients, 71 were males and 79 females. The patients ages ranged from 44 to 72 with an average of 58.

Non-tumorous tissues that were confirmed histologically were used. The relevant clinical data of patients in the study by Schröder *et al.* were age, sex, medical diagnosis (like primary or secondary liver tumour and other clinically relevant medical diagnoses), medication taken before surgery (regular drug treatment before surgery vs no drugs before surgery), cholestatic liver injury status based on liver function tests, alcohol consumption, and smoking habits. For the purpose of this study, only age and sex were used as covariates.

2.2 GENOME-WIDE GENOTYPING

318 237 SNPs were generated from genomic DNA using Illumina's HumanHap300 BeadChip. This SNP array displays a comprehensive genomic coverage across the genome of CEU from the HapMap project. In Caucasians, on average, there is one common SNP every 5.5kb across the genome. Genotype data processed as mentioned in Section 2.4 was accessed through

the accession number GSE39036 on NCBI's GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse39036).

2.3 TRANSCRIPTOME ANALYSIS

Very briefly, in the study by Schröder *et al.*, 2013, RNA was isolated from liver tissues. Agilent Bioanalyzer was used to assess high-quality RNA preparations according to Agilent RNA Integrity Number (RIN) score assignment. RIN scores range from 1 to 10. A RIN score of 1 is of the lowest quality and a RIN score of 10 is of the highest quality. RNA preparations with a RIN score greater than 7 were used. Total RNA was amplified and then labelled. cRNA quality and expression levels of more than 48 000 mRNA transcripts were assessed (Schröder *et al.*, 2013). Gene expression data was processed as described in Section 2.4 was used for this study. It was accessed through the accession number GSE32504 on NCBI's GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32504).

2.4 DATA PRE-PROCESSING AND QUALITY CONTROL

Data pre-processing and quality control done on the genotype and gene expression dataset by Schröder *et al.*, 2013 is briefly described here. Pre-processing of raw data from the HumanHap300 Genotyping BeadChip was performed. The Markov Chain Haplotyping (MaCH) imputation algorithm was used for estimating missing genotypes. As a result, 15 235 SNPs were removed from the dataset due to an extremely low call rate of less than 95%. 3668 SNPs were further excluded from the analysis. This included 3466 SNPs with minor allele frequencies (MAF) less than 4% and 201 SNPs that deviated from Hardy–Weinberg equilibrium (HWE) (false discovery rate $\leqslant 0.2$). Finally, due to a >95% genotype identity between two samples, one was excluded.

Principal component analysis (PCA) was done to detect population sub-structure amongst the samples. This did not reveal evidence for any underlying population sub-structure. As an additional quality measure, genotyping of amelogenin gene variants (which are sex chromosome-specific) and analysis of its gene expression was done. Concordance was found between observed gene expression patterns and clinically documented sex, and this was also in agreement with clinical patient documentation.

After a low-level pre-processing, 9875 genes were removed from the dataset due to >10% missing values (p-values >0.1). The study then estimated missing signal intensities. 15 439 unique genes were identified by mapping the 48 701 signal intensities.

The final processed data set consisted of 149 liver samples (71 males and 78 females) with 299 352 SNPs and 15 439 gene expression transcripts. Once the processed data was downloaded, a set of pre-processing steps were performed to develop a good quality dataset for downstream analyses.

2.5 TOOLS, SOFTWARE AND PLATFORMS USED

The following tools, software and platforms were used in this study.

2.5.1 Nextflow

Nextflow is a Groovy-based domain-specific language (DSL). It is a data-driven framework for computational pipelines that simplifies the writing of pipelines. It was developed and designed for bioinformaticians with an in-depth programming knowledge to help them solve the problem of reproducibility in data analyses. These challenges arise due to use of different computational platforms, software and database management approaches, complexity of the pipeline, and lack of good programming practice (Di Tommaso *et al.*, 2017).

Nextflow does not support a graphical user interface (GUI). It allows users to use any programming language they prefer (Di Tommaso *et al.*, 2017). It has several features that promote workflow reproducibility and portability. As a workflow management system, Nextflow can handle input and output as channels between processes and thereby reduce the need of having to create intermediate directories to store intermediate results. Variables can also be declared dynamically with no need to explicitly name files, and only the output that is required can be saved to files in each analysis step. The pipeline for this study was developed using the Nextflow version 20.04.1 framework (http://www.nextflow.io/).

2.5.2 BASH scripting

BASH scripting was used in the pipeline as it allows for consecutive commands to be executed automatically instead of one-by-one. BASH is an acronym for '**B**ourne-**A**gain **SH**ell'. It is a command language interpreter that is widely available on different operating systems and processes shell commands. BASH is the default command interpreter on most GNU/Linux systems. The Shell in BASH refers to a macro processor that allows for interactive and/or non-interactive command sessions. BASH scripting was used in this study to manipulate data.

2.5.3 PERL

PERL (https://www.perl.org) is an acronym for '**P**ractical **E**xtraction and **R**eporting **L**anguage". It is a high-level, interpreted, dynamic programming language that was originally developed for text manipulation. It was used to create the pipeline because Perl takes the best features from other languages, such as C, awk, sed, sh, and BASIC, among others and is very easy to use. PERL supports both procedural and object-oriented programming. It can interface with external C/C++ libraries. PERL is extremely portable; it can run on any operating system that has Perl interpreter installed and hence is platform independent. PERL is also free to use. PERL version 5.22.1 was used for complex data manipulation that could not be done in BASH.

2.5.4 Computing in R

R is an interpreted programming language predominantly used for statistical computing and graphics generation (https://www.r-project.org/). It is a free software environment supported by the R Foundation for Statistical Computing. R is extensively used for data analysis and is therefore popular amongst Bioinformaticians. R can be accessed by command line interface, or by RStudio which is an integrated development environment. R version 3.6.2 was used in this study to do QC on the gene expression dataset.

2.5.5 PLINK

PLINK (http://zzz.bwh.harvard.edu/plink/) is a free, open-source whole genome association analysis toolset. It is a command line tool that was designed to perform a range of basic, large-

scale data analyses in a computationally efficient manner. PLINK only works on genotype and phenotype data. In the pipeline, it was used to do QC on the genotype dataset. It does not support steps prior to data QC, like study design, planning, generating genotype or CNV calls from raw data. PLINK version 1.07 was used to do a very basic QC of the genotype dataset.

2.5.6 Genesis

Genesis is an in-house SBIMB program that is used to make admixture and PCA plots (http://www.bioinf.wits.ac.za/software/genesis/). It was used to generate publication quality pictures of population PCA and admixture charts. It is a user interactive program wherein the user can change the charts to make them more coherent and illustrative. Though other tools like R can be used to generate high-quality admixture and PCA plots, they require more expertise to use. Genesis version 0.2.5 was used to make PCA plots and to subsequently determine which samples had to be removed from the analysis because they were clear outliers.

2.5.7 Michigan Imputation Server

The Michigan Imputation Server was used for imputing into the publicly available genotype dataset (https://imputationserver.sph.umich.edu/index.html#!). It is a free Next-Generation genotype imputation service supported by the U.S. National Institutes of Health. A wide range of reference panels like HapMap Release 2, 1000 Genomes Phase 1, 1000 Genomes Phase 3, CAAPA African-American, Haplotype Reference Consortium are supported by the server. In order to impute genotype data, the dataset is uploaded to the server, the suitable reference panel is selected for the study (similar ethnic background) and the results downloaded.

2.5.8 LiftOver

LiftOver is an online tool developed by the University of California Santa Cruz that converts genome coordinates and genome annotation files between assemblies (https://genome.ucsc.edu/cgi-bin/hgLiftOver). It was used to convert genomic coordinates/ genomic position from one genome assembly to another genome assembly. It can also be used to convert dbSNP rs numbers from one build to another. It can also do both the above steps in one go, i.e. convert both genome position and dbSNP rs number over different

versions. LiftOver supports forward/reverse conversions, batch conversions, and conversions between species.

2.5.9 SNPnexus

SNPnexus is a freely available online variant annotation tool that helps in the selection and prioritisation of existing and novel genomic variants as targets for disease research and large-scale genotyping projects (https://www.snp-nexus.org/v4/) (Dayem Ullah *et al.*, 2012). It does not have to be installed and registration is not needed prior to use. After submitting a query, users are notified of the current status of the query. They can visit the result page any time to check the query status. Once analysis is completed results are made accessible on the results page; results are available for 72 hours. If a valid email address is provided by the user, query status information is sent via email. It was used to identify rsIDs associated with the identified eQTLs.

2.5.10 GTEx database

GTEx database, a US National Institutes of Health (NIH) funded project, is a resource database with an associated tissue bank for the scientific community to study and understand the relationship between genetic variation and gene expression in human tissues (https://www.gtexportal.org/home/) (Lonsdale *et al.*, 2013). It is an ongoing effort that aspires to build a comprehensive public resource to study tissue-specific gene expression and regulation. So far, samples have been collected for molecular assays like WGS, whole exome sequencing (WES), and RNA-Seq from 54 non-diseased tissue sites from almost 1000 cadavers. The GTEx Portal provides open access to data like gene expression, eQTLs, and histology images and was used to validate the novelty of the eQTLs identified in this study.

2.5.11 PharmaADME

The PharmaADME consortium is an industry-initiated effort to develop a consensus list of the known ADME genes and the likely functional variation present in them (http://pharmaadme.org/joomla/). The most important ADME genes can be grouped into four categories: Phase I and II metabolism enzymes, transporters, and modifiers. The goal of

the consortium was to develop a list of genes and genetic biomarkers that can be screened to identify predictors of pharmacokinetic variability that could impact drug safety and efficacy in the current drug development process. The list of core and extended ADME genes used in this study was obtained from PharmaADME consortium.

2.5.12 PharmGKB

The Pharmacogenomics Knowledge Base, PharmGKB, is a web-based interactive tool for researchers investigating how genetic variation affects drug response (https://www.pharmgkb.org/) (Thorn *et al.*, 2013). PharmGKB displays genetic, molecular, and clinical knowledge integrated into pathway representations. It gives Very Important Pharmacogene (VIP) summaries with links to additional external resources. Users can search and browse the knowledgebase by genes, variants, drugs, diseases, and pathways. Registration is free for the research community. Registered users can access and download data to aid in the design of future pharmacogenetics and pharmacogenomics studies. PharmGKB was used to identify drug label and variant annotations for all the ADME genes for which eQTLs were found in this study.

2.5.13 DrugBank

DrugBank is a comprehensive, free-to-access, online bioinformatics and a cheminformatics resource. It is an online database with information on drugs and drug targets (https://go.drugbank.com/) (Wishart *et al.*, 2018). Detailed drug data (chemical, pharmacological and pharmaceutical) with comprehensive drug target information (i.e., sequence, structure, and pathway) is available in DrugBank. DrugBank was used to understand how the drugs mentioned in this study were metabolised.

2.5.14 RegulomeDB

RegulomeDB is a database used to identify DNA features and regulatory elements in non-coding regions or intergenic regions of the human genome (https://regulomedb.org/) (Boyle *et al.*, 2012). The various known and predicted DNA regulatory elements include regions of DNase hypersensitivity, transcription factor binding sites, and promoter regions that have

been biochemically characterized to regulate transcription. Information on DNA features and regulatory elements have been sourced from GEO, the ENCODE project, and published literature. RegulomeDB was used to annotate the eQTLs of the core ADME genes.

2.5.15 QTLbase

QTLbase is an integrative online resource for quantitative trait loci (QTL) across multiple human molecular phenotypes (http://mulinlab.org/qtlbase/) (Zheng *et al.*, 2020). QTLbase provides an online interface that allows for phenome and tissue-wise visualisation. It annotates QTLs by integrating genomic features and functional predictions and has genome-wide QTL summary statistics for 13 human molecular traits (from 233 independent studies). QTLbase was used to validate the eQTLs found in this study.

2.5.16 Oxford Brain Imaging Genetic (BIG) Server

This browser (http://big.stats.ox.ac.uk/) supports the interpretation of GWAS results for brain imaging phenotypes and possible links to GWAS of other non-brain imaging phenotypes. It is based on data from several resources like GWAS results of ~2,000 phenotypes in the United Kingdom (UK) Biobank and results from GWAS studies like the amyotrophic lateral sclerosis GWAS, Alzheimer's disease GWAS, attention deficit hyperactivity disorder GWAS, nuclear magnetic resonance GWAS, body mass index, waist circumference and waist/hip ratio GWAS to name a few. This browser was used to identify whether any of eQTLs of the core ADME genes identified in this study were associated with drug response-related phenotypes in the UK Biobank.

2.5.17 Ensembl Variant Effect Predictor (VEP)

The Ensembl Variant Effect Predictor is a powerful tool that has both a web interface and a command-line interface. This tool can be used for the analysis, annotation, and prioritization of genomic variants in coding and non-coding regions (McLaren *et al.*, 2016). It is open source and free to use. VEP supports full reproducibility of results. The web interface of VEP release 104 was used to determine the effect of the SNPs identified as eQTLs (https://www.ensembl.org/info/docs/tools/vep/index.html).

2.6 DEVELOPING THE PIPELINE

A pipeline that will identify eQTLs will be a useful resource for researchers to identify eQTLs associated with any tissues. Such a pipeline was designed as part of this research study which incorporates a series of algorithms and filtering steps. Best practice criteria and standard operating procedures (SOP) were developed for the tools and methods used for the identification of eQTLs. The framework for the pipeline is in Nextflow and was built by Dr Phelelani Mpangase. PERL, BASH and R programming were used extensively within the pipeline. All the individual components that constitute the pipeline were written by the candidate in the above-mentioned programming languages (https://github.com/JennyMMathew/eQTL-pipeline). A basic overview of the pipeline is shown in Figure 2.1.

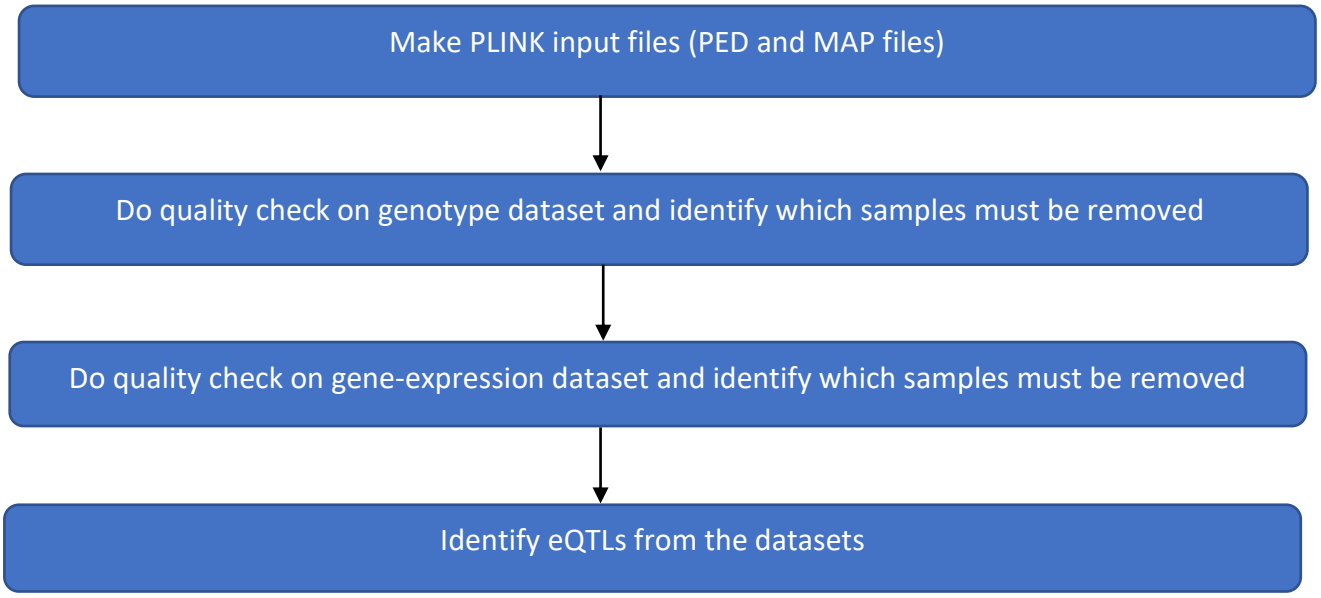


Figure 2.1: Basic overview of the pipeline. The four main functionalities of the pipeline

2.6.1 Input files or the pipeline

The genotype and gene expression datasets have three types of files available for downloading on the NCBI GEO page. They are SOFT formatted family file(s), MINiML formatted family file(s) and Series Matrix File(s). For the purpose of this study, the Series Matrix File(s) were used as input for both the genotype and gene expression dataset. Along

with the Series Matrix File(s) the platform specific supplementary files were also downloaded and used as input for the pipeline. Thus, the following four files were downloaded from NCBI GEO and used as input for the pipeline:

i. Genotype dataset (Series Matrix file)
ii. Gene expression dataset (Series Matrix file)
iii. Genotype Platform specific supplementary file
iv. Gene expression Platform specific supplementary file

2.6.2 Make PLINK standard input files (PED and MAP files)

The first step in the pipeline is to create a series of input files for PLINK. PED and MAP files are standard PLINK files that must be generated in-order to do QC of the dataset. The PLINK binary files BED, BIM and FAM are created from the PED and MAP files. Genotype QC in PLINK can only be done on the binary files.

PED and MAP files were created by a combination of PERL and BASH scripting. BED, BIM and FAM files were created from the PED and MAP files using PLINK.

2.6.2.1 PED file

PED files are PLINK pedigree files. They are space or tab delimited files. The PED file describes genetic data and the individuals in the dataset. The following six columns are mandatory in a PED file:

i. Family ID
ii. Individual ID
iii. Paternal ID
iv. Maternal ID
v. Sex
vi. Phenotype

Family ID, Individual ID, Paternal ID and Maternal ID are alphanumeric characters. The unique and specific combination of family and individual IDs helps identify a sample. The individual ID must never be 0. Paternal ID and maternal ID can be assigned 0 if the father or mother is not present in the dataset. In the fifth column, males are denoted by 1 and females by 2. Any value other than 1 or 2 refers to unknown sex. Each row in the PED file corresponds to an individual. This file should have M lines and 2N+6 columns, where M and N are the numbers of individuals and SNPs contained in the dataset respectively.

2.6.2.2 MAP file

A standard PLINK MAP file has N lines and 4 columns. N is the number of SNPs contained in the dataset. The 4 columns in a MAP file are:

i. Chromosome number (1-22, X, Y, XY, MT or 0 if unplaced)
ii. rs ID or variant identifier
iii. Genetic distance (in Morgans or centiMorgans)
iv. Base-pair position (in bp units)

Each line in the MAP file describes a single marker. Each SNP/marker must have a unique physical position. The order of the markers in the MAP file must correspond to the order of the markers in the PED file.

2.6.3 Make PLINK binary input files (BED, BIM and FAM files)

Once the PLINK standard files, the MAP and PED files, were created, they were converted into PLINK binary files BED, BIM and FAM.

2.6.3.1 BED file

The BED file contains genotype data that is represented as genotype calls of biallelic variants. This file cannot be opened and is not readable. It is encoded in binary format and is basically a binary PED (BED) file.

2.6.3.2 BIM file

The BIM file is a variant information file. It is an extended version of the MAP file. It has the following columns:

i. Chromosome number (1-22, X, Y, XY, MT or 0 if unplaced)
ii. rs id or variant identifier
iii. Genetic distance (in Morgans or centiMorgans)
iv. Base-pair position (in bp units)
v. Allele 1 (usually the minor allele)
vi. Allele 2 (usually the major allele)

2.6.3.3 FAM file

This file contains Sample information and has the following columns:

i. Family ID
ii. Individual ID
iii. Paternal ID
iv. Maternal ID
v. Sex
vi. Phenotype

The FAM file is essentially just the first six columns of the PED file.

2.6.4 Quality check on genotype dataset

A thorough assessment of genotype data quality was carried out in-order remove sub-standard samples and genetic markers (SNPs) over several steps. Quality check can be further sub-divided into sample QC (2.6.4.1-2.6.4.5) and SNP QC (2.6.4.6-2.6.4.8). Standard cut-offs were used in SNP QC (Anderson *et al.*, 2010) except for identification of SNPs of high missingness rate. In this pipeline, SNP QC was followed by sample QC.

2.6.4.1 Identification of individuals with discordant sex information

The very first step was to check whether there were any discrepancies between information on sex from the genotype data and the recorded sex of the samples in the phenotype data.

This step checked that the genetic data was consistent with the sex mentioned in the FAM or PED files. Sex check was carried out by calculating the homozygosity rate across all SNPs from the X chromosome for each individual and comparing it to the expected rate. The expected homozygosity rate for males is 1 and for females is < 0.2. A sample is identified as a problem when the sex provided with the data does not correspond to the sex as determined by the X chromosome data. Such samples were excluded from further analysis.

2.6.4.2 Identification of individuals with high missingness rate

The quality of DNA influences both genotype call rate and accuracy. Poor DNA quality or low DNA concentration can lead to low call rates and genotype accuracy. In this step, samples with very poor DNA quality were removed from further analysis. Depending on what threshold is set, this step will keep all individuals that have at least (100-specified call rate). Individuals with more than 5% missingness were removed (cut off 0.05).

2.6.4.3 Identification of individuals with extreme mean heterozygosity

All individuals have a certain percentage of heterozygous genotypes. Excessive mean heterozygosity indicates possible DNA sample contamination, whereas reduced mean heterozygosity points towards inbreeding. Thus, individuals with very high or extremely low proportion of heterozygous genotypes were removed from further analysis. The threshold for extreme heterozygosity was determined by generating a plot of observed heterozygosity rate per individual (y-axis) and missing SNPs per individual (x-axis). An observed heterozygosity rate was calculated from the samples in the dataset and then the generated plot was used to decide an appropriate threshold for the removal of individuals with extreme mean heterozygosity. A common threshold for inclusion/exclusion is a heterozygosity rate ± 3 standard deviations from the mean.

2.6.4.4 Identification of related or duplicate individuals

Samples in duplicate or individuals that are second-degree relatives or higher can lead to an over-representation of genotypes and will introduce a bias in the analysis. Hence, ideally, we want all individuals in the study to be unrelated. Thus, in this step, duplicated or related samples were removed. An identity by descent (IBD) score is used to identify related samples as IBD is the degree of recent shared ancestry between a pair of individuals. In PLINK, an IBD

score is denoted as PI_HAT. In each pair of individuals with IBD score of PI_HAT > 0.1875, the individual in the pair with the lowest call rate was removed.

2.6.4.5 Identification of population outliers

Population stratification in a dataset could lead to the identification of genotypic differences that may be due to different population origins rather than to the parameters that the study was seeking to identify. Hence population outliers were removed to reduce the effect of population stratification because it can be a potential source of bias in the analysis.

The dataset used for this study was primarily a dataset from individuals of European origin and described relative to the NCBI36/hg18 build. A PC plot was generated with the dataset to identify genetic outliers. Two PC plots were then generated with the dataset and three other populations in order to ensure that none of the samples in the dataset had significant admixture with the three other populations. The three other populations used were from the 1000 Genomes Project and they are the following:

i. CEU (Utah Residents with Northern and Western European Ancestry from the CEPH collection)
ii. ASW (Americans of African Ancestry in Southwest USA – this is an admixed population with some African ancestry), and
iii. YRI (Yoruba in Ibadan, Nigeria)

The super population code for CEU is EUR (European) and that of ASW and YRI is AFR (African). Since the dataset used for this study is from individuals of European origin, samples were expected to cluster with CEU. ASW and YRI were used to identify whether the dataset had any African admixture. The three populations used from the 1000 Genomes Project were from a different build, GRCh37/hg19. Since the dataset used for the study and the three populations from 1000 Genomes Project were in different builds, LiftOver was used to convert the genotype dataset to hg19 build. Samples that did not cluster with the majority of the dataset and those that clustered with ASW and YRI were excluded from further analysis, as GSE390036 is a dataset from individuals of European origin.

2.6.4.6 Identification of SNPs with high missingness rate

All SNPs with a call rate less than 93.5% were removed in this step. SNPs with more than 6.5% missing data (threshold = 0.065) were removed in order to increase stringency, by reducing random variability (noise) from data without losing too many SNPs.

2.6.4.7 Identification of SNPs with low minor allele frequency (MAF)

Minor allele frequency (MAF) is the frequency of the rarer allele in a population. It differentiates common SNPs from rare SNPs in a population. The threshold used in this study for this step is 1%. This step thus removed all SNPs with a MAF that was lower than 1% (threshold= 0.01).

2.6.4.8 Identification of SNPs deviating from Hardy-Weinberg Equilibrium (HWE)

The Hardy-Weinberg Equilibrium principle states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. Evolutionary influences can be mutation, selection, genetic drift to name a few. It is good practice to remove SNPs from the dataset that deviate significantly from HWE because deviations can be because of genotyping error, genotype-calling error, inbreeding or genetic drift, population bottlenecks or founder events. Hence this step removed all SNPs that deviated from the Hardy-Weinberg equilibrium with a p-value cut off of 0.00001.

From all the above genotype QC steps, samples and SNPs that needed to be excluded from further analysis were identified and excluded.

2.6.5 Quality check on Gene Expression dataset

A basic gene expression normalization and quality control was done on the gene expression dataset. This was done using R. The available gene expression values were log2-transformed. Then a PCA was performed to detect potential outliers in the dataset. Since all the samples in the dataset were from the liver, they were expected to cluster together. Samples that did not cluster with the majority of the dataset were removed from further analysis. Dr Stanford Kwenda graciously helped perform quality check on the gene expression dataset.

2.7 IMPUTATION

Prior to doing imputation, the dataset's build had to be changed from hg18 to hg19. This was done using the online tool LiftOver. Since the genome build was changed to hg19, a very basic QC was done again using standard cut-offs except for the removal of SNPs with a call rate of less than 95%. The QC steps done prior to imputation use were:

i. All individuals with a missingness of higher than 95% were excluded from further analysis.
ii. All SNPs with a call rate less than 95% were removed.
iii. All SNPs with a MAF that is lower than 1% were removed from further analysis.
iv. All SNPs that deviate from the HWE threshold of 0.00001 were removed.

My colleague Dr Dhriti Sengupta performed the imputation using the Michigan imputation server. The Phase 3 1000 Genomes Project reference panel was used.

Post-imputation, QC was done on the imputed dataset. The cut-offs used are standard and will reduce 'noise' from data without losing too many SNPs (Danecek *et al.*, 2011). It included the following steps:

i. r^2 value (imputation score) must be greater than 0.8 (r^2 shows how reliable the imputed data is. A score of one means that data has been imputed with full confidence.)
ii. Lower minor allele count bound must be greater than one (mac > 1). (Allele count is the number of times an allele is observed in a dataset. This step only included sites with minor allele count greater than one.)
iii. SNPs had to have a minimum of two alleles (min-alleles 2) and a maximum of two (max-alleles 2). (This step only retained sites with number of alleles is equal to two.)
iv. SNPs with missing data of up to 60% were tolerated (max-missing 0.4). (This step excluded SNPs based on the proportion of missing data.)

Checks on minor allele frequency (MAF) and departure from Hardy-Weinberg equilibrium (HWE) post-imputation were not done as this dataset is relatively small. Post-imputation checks on MAF and HWE is only necessary when working with large datasets. Moreover, a MAF cut-off, albeit a lower one, is being applied indirectly by setting mac> 1 post-imputation.

The imputed dataset was then used to identify eQTLs using the Matrix eQTL package. Prior to running the R package, the genome build of the coordinates in the SNP location file and gene location file were also converted to hg19 using LiftOver.

2.7.1 Identification of eQTLS from the genotype and gene expression dataset

eQTL analysis links genotype to variations in gene expression levels. Since eQTL analysis involves testing for associations between billions of transcript-SNP pairs, it is a computationally intensive task. A package in R, namely, the MatrixEQTL package was used to identify eQTLs because it is found to be almost three-fold faster than existing popular tools for eQTL analysis.

MatrixEQTL package needs five different types of files as input (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/Sample_Data/). Measurements in all these files must be in numeric. The five input files needed are:

1. Covariate file
2. Genotype file
3. Gene expression file
4. Gene location file
5. SNP location file

The covariate, genotype and gene expression files must have the same number of columns. If a dataset has N samples, its corresponding covariate, genotype and gene expression files will have N+1 columns. The additional column in all three files is the very first column, the row header which lists out which covariate/SNP/gene information is listed in that row. The columns of all these three files must be matching in order. The Nth column in all these files

will contain information about the (N-1)th sample. For example, the tenth column in the covariate, genotype and gene expression file will have covariate/genotype/gene expression information corresponding to the ninth sample in the dataset. Examples of the covariate, genotype and gene expression files are shown in Figure 2.2, Figure 2.3 and Figure 2.4.

id	Sam_01	Sam_02	Sam_03	Sam_04	Sam_05	Sam_06	Sam_07	Sam_08	Sam_09	Sam_10
gender	0	0	0	0	0	0	0	0	1	1
age	36	40	46	65	69	43	40	54	44	70

Figure 2.2: An example of a covariate file

id	Sam_01	Sam_02	Sam_03	Sam_04	Sam_05	Sam_06	Sam_07	Sam_08	Sam_09	Sam_10
Snp_01	2	0	2	0	2	1	2	1	1	1
Snp_02	0	1	1	2	2	1	0	0	0	1
Snp_03	1	0	1	0	1	1	1	1	0	1
Snp_04	0	1	2	2	2	1	1	0	0	0
Snp_05	1	1	2	1	1	2	1	1	0	1
Snp_06	2	2	2	1	1	0	1	0	2	1
Snp_07	1	1	2	2	0	1	1	1	1	0
Snp_08	1	0	1	0	1	0	0	1	1	1
Snp_09	2	1	2	2	0	1	1	0	2	1
Snp_10	1	1	0	0	0	2	2	1	1	2
Snp_11	2	2	2	0	2	1	1	2	1	2
Snp_12	1	1	2	2	2	1	1	1	1	0
Snp_13	0	1	1	1	1	1	2	1	2	2
Snp_14	0	0	1	0	1	2	2	2	1	1
Snp_15	1	2	1	2	2	1	2	1	1	2

Figure 2.3: An example of a genotype file

id	Sam_01	Sam_02	Sam_03	Sam_04	Sam_05	Sam_06	Sam_07	Sam_08	Sam_09	Sam_10
Gene_01	4.91	4.63	5.18	5.07	5.74	5.09	5.31	5.29	4.73	5.72
Gene_02	13.78	13.14	13.18	13.04	12.85	13.07	13.09	12.83	14.94	13.86
Gene_03	12.06	12.29	13.07	13.65	13.86	12.84	12.29	13.03	13.13	14.93
Gene_04	11.63	11.88	12.74	12.66	13.16	11.99	11.97	12.81	11.74	13.29
Gene_05	14.72	14.66	14.63	15.91	15.46	14.74	15.17	15.01	14.05	14.90
Gene_06	12.29	12.23	12.47	13.21	12.63	12.18	12.44	12.45	13.22	12.70
Gene_07	12.56	12.71	12.49	13.41	13.60	12.35	12.27	13.42	14.73	15.47
Gene_08	12.27	12.58	12.61	13.02	12.86	12.32	12.30	13.19	15.21	14.60
Gene_09	9.82	9.29	8.95	8.18	8.11	9.43	9.17	9.25	10.95	9.82
Gene_10	14.24	14.52	14.62	13.65	13.46	14.04	13.97	13.73	12.29	12.01

Figure 2.4: An example of a gene expression file

The gene location file must have four columns. The first column lists out all gene names, the second column shows which chromosome the gene in found in. It is listed as ‘chr’ followed by the chromosome number without any space in between the two. For example, if the gene is in chromosome 10, it must be listed as chr10. The third column has chromosome start coordinate, and the fourth column has the chromosome end coordinate. Figure 2.5 shows an example of the gene location file.

geneid	chr	s1	s2
Gene_01	chr1	721289	731289
Gene_02	chr1	752565	762565
Gene_03	chr1	777121	787121
Gene_04	chr1	785988	795988
Gene_05	chr1	792479	802479
Gene_06	chr1	798958	808958
Gene_07	chr1	888658	898658
Gene_08	chr1	918572	928572
Gene_09	chr1	926430	936430
Gene_10	chr1	1000000	1010000

Figure 2.5: An example of a gene location file

The SNP location file must have three columns. The first column of this file lists out rsIDs. The second column describes on which chromosome the SNP is found. It is listed as ‘chr’ followed by the chromosome number without any space in between the two. The third and the last column in this file lists out the chromosome coordinate of the SNP. Figure 2.6 shows an example of the gene location file.

snp	chr	pos
Snp_01	chr1	721289
Snp_02	chr1	752565
Snp_03	chr1	777121
Snp_04	chr1	785988
Snp_05	chr1	792479
Snp_06	chr1	798958
Snp_07	chr1	888658
Snp_08	chr1	918572
Snp_09	chr1	926430
Snp_10	chr1	947033
Snp_11	chr2	721289
Snp_12	chr2	752565
Snp_13	chr2	777121
Snp_14	chr2	785988
Snp_15	chr2	792479

Figure 2.6: An example of a SNP location file

A detailed overview of the pipeline is shown in Figure 2.7 and Figure 2.8. The genotype series matrix file and platform specific supplementary file were used as input to create the PED and MAP files. Four PERL scripts were used with basic BASH commands to extract information from the input files and to manipulate them into the required format of the PED file. A single PERL script was used to manipulate data to make the MAP file. PED and MAP files were then used as input in PLINK to generate PLINK binary files, the BED, BIM and BAM files. BED, BIM and BAM files were used to do QC on the genotype dataset. QC identified which samples needed to be excluded from further analysis and which SNPs met all the QC criteria.

QC on the gene expression dataset was done by plotting a PCA in R. The input file used for this step was the gene expression series matrix file. The outliers in the plot were samples that failed QC. These samples were removed before running the matrix eQTL R package.

The samples in the genotype and gene expression dataset used different sample identifiers. To determine which samples have failed QC in both datasets, samples from both genotype and gene expression dataset were mapped. Mapping of sample title, sample id and sample description was done to correctly identify samples that needed to be removed from further analysis. Once the samples were identified, the next part of the pipeline generated the five

input files for Matrix eQTL, namely the covariate file, the SNP location file, the genotype file, the gene expression file and the gene location file.

The covariate file was created from the genotype series matrix file. Information in the series matrix file was manipulated such that it was in the format shown in Figure 2.3. For this analysis, only age and sex are used as covariates.

The SNP location file was made from the genotype platform specific supplementary file and the file that has information on all SNPs that passed QC. Essentially, data in the genotype platform specific supplementary file was manipulated into three columns like in Figure 2.7. After data was in the required format, SNPs that failed QC were removed. SNPs in this file were also listed in the exact same order as SNPs in the genotype file.

Genotype series matrix file was used as input for creating the genotype file along with the file that has information on SNPs that have failed QC. Data in the input file was first manipulated to be in the format shown in Figure 2.4. Once it was in that format, all genotype information on SNPs that have failed QC was removed.

The next step was to make the gene expression file. The gene expression series matrix file and the gene expression platform specific supplementary file were used as input. Once data from the series matrix file was manipulated to be in the format shown in Figure 2.5, the platform specific supplementary file was used to replace Illumina ID's with gene names.

Only the gene expression platform specific supplementary file was used as input to create the gene location file. Once all empty probe coordinates were removed from the input file, all remaining probe coordinates were captured. Then all coordinates with _ or | in the chromosome number were removed and file is in the format shown in Figure 2.6. Once all five input files needed to run Matrix eQTL package were created, the samples that failed

genotype and gene expression QC and were mapped earlier in the pipeline were removed from the covariate, genotype and gene expression files. The R package Matrix eQTL was then run and the list of *cis*- and *trans*-eQTLs are generated. *Cis*-eQTLs or local eQTLs are detected within 1 megabase (Mb) on either side of a gene's transcription start site. *Trans* or distant eQTLs are detected at least 5 Mb downstream or upstream of the TSS or on a different chromosome. Though this pipeline generates both *cis*- and *trans*-eQTLs, in this study, the focus is on *cis*-eQTLS. Only these will be described further, as *trans*- eQTLs are often not reproduced and validated across different studies.

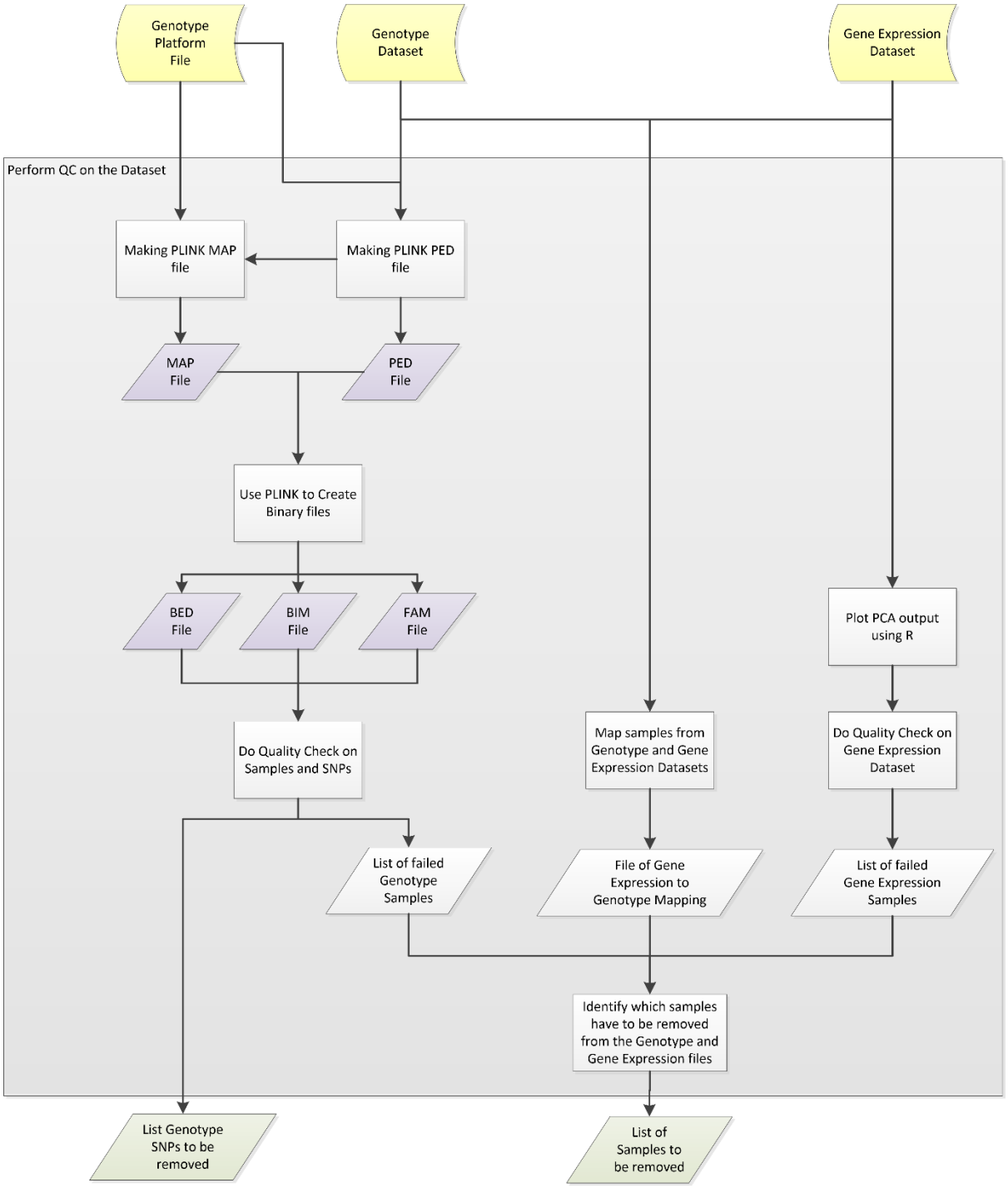


Figure 2.7: Pipeline - part one. Detailed QC for genotype and gene expression dataset is described here.

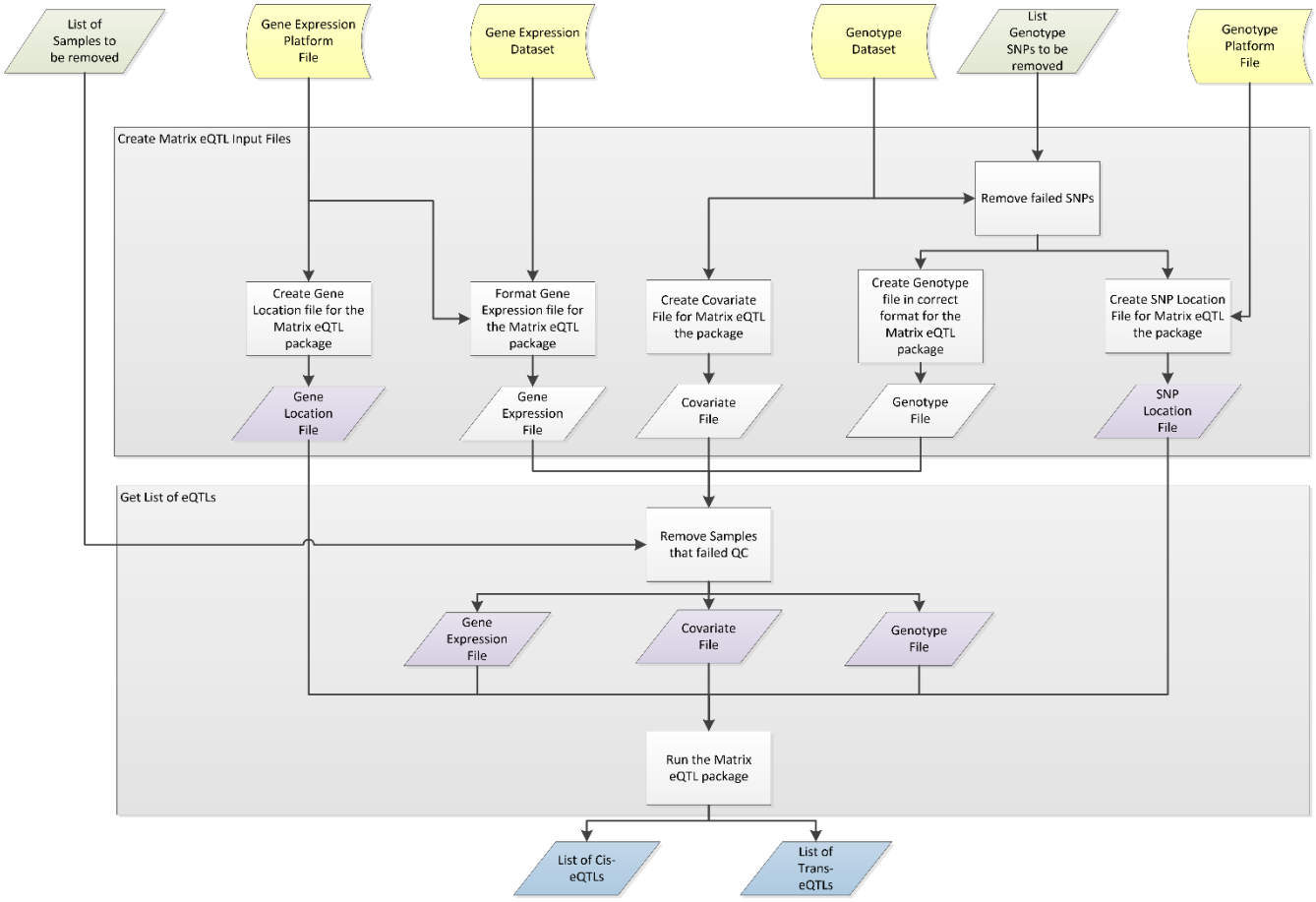


Figure 2.8: Pipeline - part two. This shows how input files needed to run Matrix eQTL are created and how *cis*- and *trans*-eQTLs are obtained.

2.8 IDENTIFICATION OF eQTLs

The pipeline can be used for studying a wide range of genes and tissues, however, in this study it is applied to liver transcriptome data and a sub-set of ADME genes, as the focus is on eQTLs of ADME genes.

The PharmaADME consortium http://pharmaadme.org/joomla/) has two comprehensive lists of ADME genes that are referred to in detail. They are:

i. Core ADME gene list: The genetic markers in this list are standardized and evidence based. The 32 core ADME genes are listed in

ii. Table 2.1.

iii. Extended ADME gene list: This list has 267 ADME genes associated with drug metabolism. They are listed in Table 2.2.

Table 2.1: Core ADME gene list

Gene name	Chr number	Gene start (bp)	Gene end (bp)
CYP2C19	10	94762681	94855547
CYP2C9	10	94938658	94990091
CYP2C8	10	95036772	95069497
ABCC2	10	99782640	99852594
CYP2E1	10	133520406	133561220
SLC22A6	11	62936385	62984967
GSTP1	11	67583742	67586656
SLCO1B3	12	20810702	20916911
SLCO1B1	12	21131194	21239796
CYP1A1	15	74719542	74725536
CYP1A2	15	74748845	74756607
SULT1A1	16	28605196	28623625
CYP2A6	19	40843541	40850447
CYP2B6	19	40991282	41018398
UGT1A1	2	233760270	233773300
CYP2D6	22	42126499	42130881
SLC15A2	3	121894401	121944188
UGT2B17	4	68537184	68568527
UGT2B15	4	68646597	68670652
UGT2B7	4	69051363	69112987
ABCG2	4	88090150	88231628
TPMT	6	18128311	18155077
SLC22A1	6	160121815	160158718
SLC22A2	6	160171061	160277638
ABCB1	7	87503017	87713323
CYP3A5	7	99648194	99679998
CYP3A4	7	99756960	99784248
NAT1	8	18170477	18223689
NAT2	8	18391282	18401218
GSTM1	1	109687814	109709039
DPYD	1	97077743	97995000

Table 2.2: Extended ADME gene list

Extended ADME genes						
ABCB8	*UGT2B11*	*CBR3*	*PNMT*	*SLCO2A1*	*GSS*	*DDO*
ABCC12	*UGT2B28*	*CDA*	*PON1*	*SLCO3A1*	*GSTCD*	*DHRS1*
ABCC3	*UGT2B4*	*CYB5R3*	*PON2*	*SLCO4A1*	*HNF4A*	*DHRS12*

ABCC4	*ABCA1*	*CYP11A1*	*PON3*	*SLCO4C1*	*HNMT*	*DHRS3*
AHR	*ABCA4*	*CYP11B1*	*POR*	*SLCO5A1*	*HSD11B1*	*DHRS4*
ALDH4A1	*ABCB11*	*CYP11B2*	*PPARD*	*SLCO6A1*	*HSD17B11*	*DHRS4L1*
ALDH5A1	*ABCB4*	*CYP17A1*	*PPARG*	*SULT1C1*	*HSD17B14*	*DHRS4L2*
ALDH6A1	*ABCB5*	*CYP1B1*	*RXRA*	*SULT1C2*	*LOC731356*	*DHRS7*
CES1	*ABCB6*	*CYP20A1*	*SLC10A2*	*SULT1E1*	*MGST1*	*DHRS7B*
CES2	*ABCB7*	*CYP21A2*	*SLC13A1*	*SULT2A1*	*MGST2*	*DHRS7C*
CYP7A1	*ABCC1*	*CYP24A1*	*SLC13A2*	*SULT2B1*	*MGST3*	*DHRS9*
EPHX1	*ABCC10*	*CYP26A1*	*SLC13A3*	*TAP1*	*MPO*	*DHRSX*
FMO3	*ABCC11*	*CYP27A1*	*SLC16A1*	*UGT1A10*	*NOS1*	*DPEP1*
GSTA1	*ABCC5*	*CYP2A13*	*SLC19A1*	*UGT1A4*	*NOS2A*	*FMO6P*
GSTA2	*ABCC6*	*CYP2A7*	*SLC22A10*	*UGT1A5*	*NOS3*	*HAGH*
GSTA3	*ABCC8*	*CYP2C18*	*SLC22A12*	*UGT2B10*	*PPARA*	*IAPP*
GSTA4	*ABCC9*	*CYP2F1*	*SLC22A13*	*ABCC13*	*SERPINA7*	*KCNJ11*
GSTA5	*ABCG1*	*CYP2J2*	*SLC22A14*	*ARSA*	*SLC7A7*	*LOC728667*
GSTM2	*ADH1A*	*CYP39A1*	*SLC22A15*	*CAT*	*SOD1*	*LOC731931*
GSTM3	*ADH1B*	*CYP3A43*	*SLC22A16*	*CHST8*	*SOD2*	*MAT1A*
GSTM4	*ADH1C*	*CYP3A7*	*SLC22A17*	*CYP19A1*	*SOD3*	*METAP1*
GSTO1	*ADH4*	*CYP4B1*	*SLC22A18*	*CYP26C1*	*SULF1*	*PDE3A*
GSTO2	*ADH5*	*CYP4F11*	*SLC22A18AS*	*CYP27B1*	*SULT4A1*	*PDE3B*
GSTT2	*ADH6*	*CYP51A1*	*SLC22A3*	*CYP2R1*	*TAP2*	*PLGLB1*
SLC10A1	*ADH7*	*EPHX2*	*SLC22A4*	*CYP2S1*	*UGT8*	*ATP7A*
SLC15A1	*ALDH1A1*	*FMO1*	*SLC22A5*	*CYP46A1*	*XDH*	*ATP7B*
SLC22A11	*ALDH1A2*	*FMO2*	*SLC22A7*	*CYP4A11*	*ADHFE1*	
SLC22A8	*ALDH1A3*	*FMO4*	*SLC22A9*	*CYP4F12*	*CHST1*	
SLC7A5	*ALDH1B1*	*FMO5*	*SLC27A1*	*CYP4F2*	*CHST10*	
SLCO1A2	*ALDH2*	*GPX2*	*SLC28A1*	*CYP4F3*	*CHST11*	
SLCO2B1	*ALDH3A1*	*GPX3*	*SLC28A2*	*CYP4F8*	*CHST12*	
SULT1A2	*ALDH3A2*	*GPX7*	*SLC28A3*	*CYP4Z1*	*CHST13*	
SULT1A3	*ALDH3B1*	*GSR*	*SLC29A1*	*CYP7B1*	*CHST2*	
SULT1B1	*ALDH3B2*	*GSTK1*	*SLC29A2*	*CYP8B1*	*CHST3*	

UGT1A3	*ALDH7A1*	*GSTM5*	*SLC2A4*	*DHRS13*	*CHST4*	
UGT1A6	*ALDH8A1*	*GSTZ1*	*SLC2A5*	*DHRS2*	*CHST5*	
UGT1A7	*ALDH9A1*	*NNMT*	*SLC5A6*	*GPX1*	*CHST6*	
UGT1A8	*AOX1*	*NR1I2*	*SLC6A6*	*GPX4*	*CHST7*	
UGT1A9	*ARNT*	*NR1I3*	*SLC7A8*	*GPX5*	*CHST9*	
UGT2A1	*CBR1*	*CFTR*	*SLCO1C1*	*GPX6*	*CYP2D7P1*	

2.8.1 In-silico analyses of eQTLs

An in-depth in-silico analysis of each eQTLs identified in the above step was done using the following steps:

i. Identify whether the eQTLs have rsIDs associated with them. This was done using the online tool SNPnexus.

ii. Identify whether the ADME gene for which the eQTL has been identified had previously been shown to be expressed in the liver. Since the transcriptome dataset used in this study was obtained from liver samples, it is safe to assume that genes in this dataset are expressed in the liver. However, this step was done to independently validate whether the genes have indeed previously been shown to be expressed in the liver. This was done by searching the GTEx database.

iii. Identify whether the eQTLs exist in the GTEx database and QTLbase database. GTEx was searched for the previously identified eQTLs. GTEx eQTL Calculator was used to check whether the eQTLs identified are novel or not. It required input to be in the format: Variant ID, Gene ID, Tissue name.

The following were allowed for Variant ID, Gene ID and Tissue name:

a. Variant ID: GTEx variant ID or dbSNP rs number.
b. Gene ID: Gencode ID, Ensembl ID, or HGNC Gene Symbol.
c. Tissue name must use allowable tissue names (tissue names that are used in the GTEx database)

The QTLbase database allows users to search for 'Traits' or 'Variants'. For this study, the core ADME genes were queried under 'Traits'.

iv. Make violin plots for all the eQTLs to visualise the gene expression distribution for each of the three genotypes of the eQTL using R.

v. Determine the frequencies of the eQTLs in different African populations using different datasets, the AWI-Gen whole genome sequence (WGS) data (n=100 WGS samples of Southeastern Bantu-speakers), H3A-Baylor WGS data (n=348), 1000 Genomes Project (KGP) African populations and African Genome Variation Project (AGVP) WGS data (n=320). The effect size and the direction of effect for eQTLs are found to be highly conserved across populations. Since there is very minimal information about eQTLs in African populations, we are assuming that eQTLs identified in populations of European ancestry are also eQTLs in populations of African ancestry. There is limited or no data where this can be tested, and therefore discordance cannot be assessed.

vi. Understand the effect imputation has on the number of eQTLs identified. This was done by comparing the number of eQTLs identified pre-imputation and post-imputation.

vii. Understand the impact that the removal of samples has on the number of eQTLs identified. Steps in genotype QC include PC analysis. Outliers could therefore be excluded. A two-step approach was used for the analysis. The first analysis was to be done on all samples, except samples with high missing data rate and PC outliers. The second analysis was to be done on samples that passed both genotype and gene expression QC and PC analysis. The number of eQTLs generated with each approach was compared.

viii. Identify whether the eQTLs for core ADME genes are associated with drug response-related phenotypes in the UKBiobank (http://big.stats.ox.ac.uk/). This was achieved by querying the database using the rsIDs associated with the eQTL.

ix. Annotate the eQTLs identified for the core ADME genes using RegulomeDB and VEP. The rsIDs associated with the eQTLs were queried in the databases and the results were analysed.

x. Understand the role of the eQTLs with respect to pharmacogenetics. This was done by looking at drug label and variant annotations for all the core ADME genes for which

eQTLs were found in PharmGKB. PharmGKB has a collection of drug label annotations which includes pharmacogenetic information approved by pharmaceutical regulatory bodies of the United States of America (US FDA), Europe (EMA), Switzerland (Swiss Agency of Therapeutic Products (Swissmedic), Japan (Pharmaceuticals and Medical Devices Agency (PMDA) and Canada (Health Canada/Santé Canada (HCSC). These annotations give a summary of the pharmacogenetics of alleles of these genes with respect to metabolism of drugs. It provides critical guidance from the drug labels for patients with certain genotype/metaboliser phenotypes. Each drug label annotation provided in PharmGKB has one of four different tags (Testing required, Testing Recommended, Actionable PGx and Informative PGx). These tags indicate the level of action needed for the information generated according to each drug label.

The label 'Testing required' implies that some sort of testing (gene, protein, chromosomal, genetic testing, functional protein assays, cytogenetic studies) is required before prescribing the drug. If the drug label mentions that a test must be performed, this is interpreted as a prerequisite. 'Testing recommended' implies that some sort of testing is recommended before using this drug. This recommendation might be applicable only to a subset of patients. If the drug label mentions that genotyping or phenotyping must be done or that testing must be considered prior to drug use, PharmGKB highly recommends performing a test. The label 'Actionable PGx' gives information about changes in drug efficacy, dosage levels, metabolism and/or toxicity because of gene/protein/chromosomal variants or certain phenotypes like poor metabolisers. This label does not require or recommend any form of testing prior to commencement of drug use. However, this label might mention drug contraindications in a subgroup of patients with certain genotype/metaboliser phenotype. 'Informative PGx' contains information pertaining to genotypes/metaboliser phenotypes that do/do-not affect drug efficacy, drug dosage levels, metabolism or toxicity. This information, however, is not deemed clinically significant/actionable and therefore testing is neither recommended nor required prior to drug use. PharmGKB also gives information related to dosage levels and dosage level adjustments that must be made based on the genotype/metaboliser phenotype. Drugs that are contraindicated with certain genotype/ metaboliser

phenotypes are also mentioned in PharmGKB where applicable. Even though the majority of the recommendations were based on coding variants, the eQTLs have the potential to affect dosing recommendations as the identified eQTLs were predicted to alter the expression of the core ADME gene it is associated with.

For the extended ADME genes, a list of commonly prescribed drugs was identified by looking for drugs that treat the top ten causes of mortality in South Africa published by the Department of Statistics, South Africa in 2017. The causes of mortality that were focussed on in this study are HIV and AIDS, tuberculosis, diabetes, hypertension, and cardiovascular diseases. We have also looked at the most frequently bought over-the-counter medications and prescription medications for pain. The commonly bought over-the-counter drugs and prescription drugs for the above mentioned conditions are listed in Table 2.3, Table 2.4, Table 2.5, Table 2.6, Table 2.7, Table 2.8 and Table 2.9.

Table 2.3: Drugs commonly prescribed in South Africa for HIV and AIDS

Drugs for HIV and AIDS		
Tenofovir	Rilpivirine	Stavudine
Maraviroc	Delavirdine	Didanosine
Enfuvirtide	Ritonavir	Abacavir
Vicriviroc	Efavirenz	Emtricitabine
Raltegravir	Etravirine	Fosamprenavir
Dolutegravir	Nevirapine	Tipranavir
Elvitegravir	Zidovudine	Amprenavir
Zalcitabine	Lamivudine	Lopinavir
Atazanavir	Nelfinavir	Indinavir
Darunavir	Saquinavir	

Table 2.4: Drugs commonly prescribed in South Africa for diabetes

Drugs for Diabetes
Metformin
Pioglitazone

Table 2.5: Drugs commonly prescribed in South Africa for tuberculosis

Drugs for Tuberculosis
Rifampicin
Isoniazid
Pyrazinamide
Ethambutol
Bedaquiline
Linezolid
Pretomanid

Table 2.6: Drugs commonly prescribed in South Africa for hypertension

Drugs for Hypertension
Hydrochlorothiazide
Enalapril
Amlodipine
Spironolactone

Table 2.7: Drugs commonly prescribed in South Africa for cardiovascular disease

Drugs for Cardiovascular disease
Rosuvastatin
Atorvastatin
Simvastatin
Lovastatin
Fluvastatin
Pravastatin

Table 2.8: Drugs commonly prescribed in South Africa for cancer

Drugs for Cancer
Fluorouracil
Capecetabine
Cisplatin
Cyclophosphamide
Irinotecan

Table 2.9: Commonly used analgesics in South Africa

Commonly prescribed Analgesics
Acetaminophen
Naproxen
Tapentadol

CHAPTER 3 RESULTS

In order to run the pipeline to identify eQTLs, basic sample and SNP QC was carried out on the genotype dataset, GSE39036. The gene expression dataset, GSE32504 was already normalized, therefore only a sample level QC was done.

3.1 GENOTYPE DATA: SAMPLE-RELATED QUALITY CONTROL

The genotype dataset, GSE39036, was downloaded from NCBI's GEO and included a total of 149 samples. QC of the genotype dataset resulted in four samples being removed and excluded from further analyses because of high missingness (greater than 5%). No samples were excluded due to sex discordance. Extreme mean heterozygosity was not found in any of the samples. There were no related or duplicate samples. However, three samples were found to be population outliers and are highlighted in yellow in the PC plot (Figure 3.1). Figure 3.2 and Figure 3.3 are PC plots generated with the genotype dataset and three populations from the 1000 Genomes Project namely CEU, ASW and YRI. The samples highlighted in yellow in Figure 3.1, Figure 3.2 and Figure 3.3 were the same samples and were excluded from further analysis as they did not cluster together with other samples from the dataset. If outliers are not excluded, they may confound the results because of population stratification. The sample highlighted in pink, GSM954341, is also an outlier (Figure 3.1). Unfortunately, this was only spotted by the examiner and the decision was made not to repeat all the analyses without this sample, as the deletion of a single sample is unlikely to significantly influence the outcome.

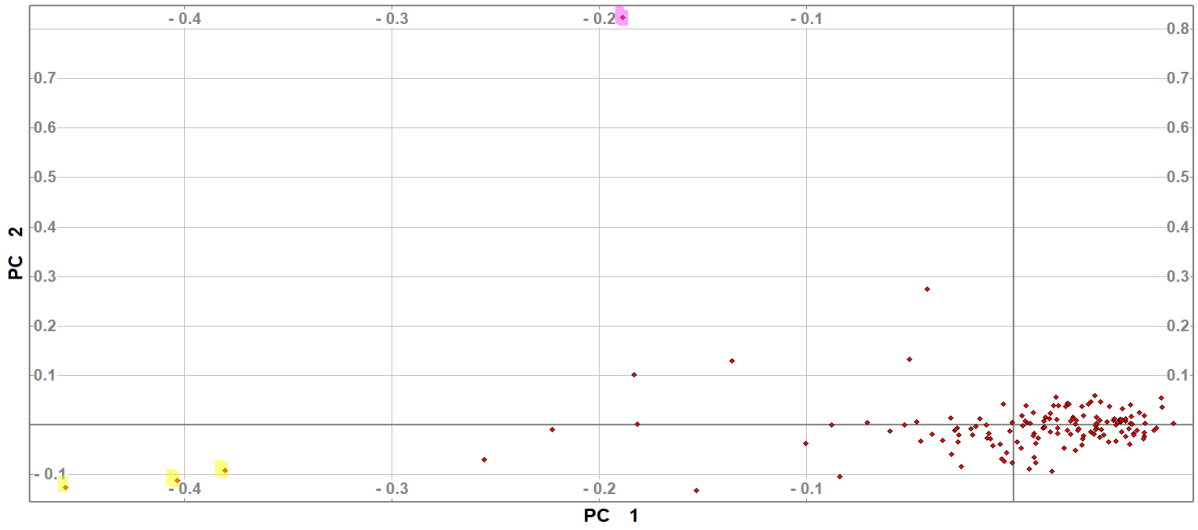


Figure 3.1: Clustering of samples using genotyping data. PC plot for 149 samples using 299 352 SNPs from the genotyping data revealed three outliers (shown in yellow) that were removed from further analysis. The sample shown in pink was not excluded as it clustered with GSE (samples from the dataset used in this study) and CEU in Figure 3.2 and 3.3.

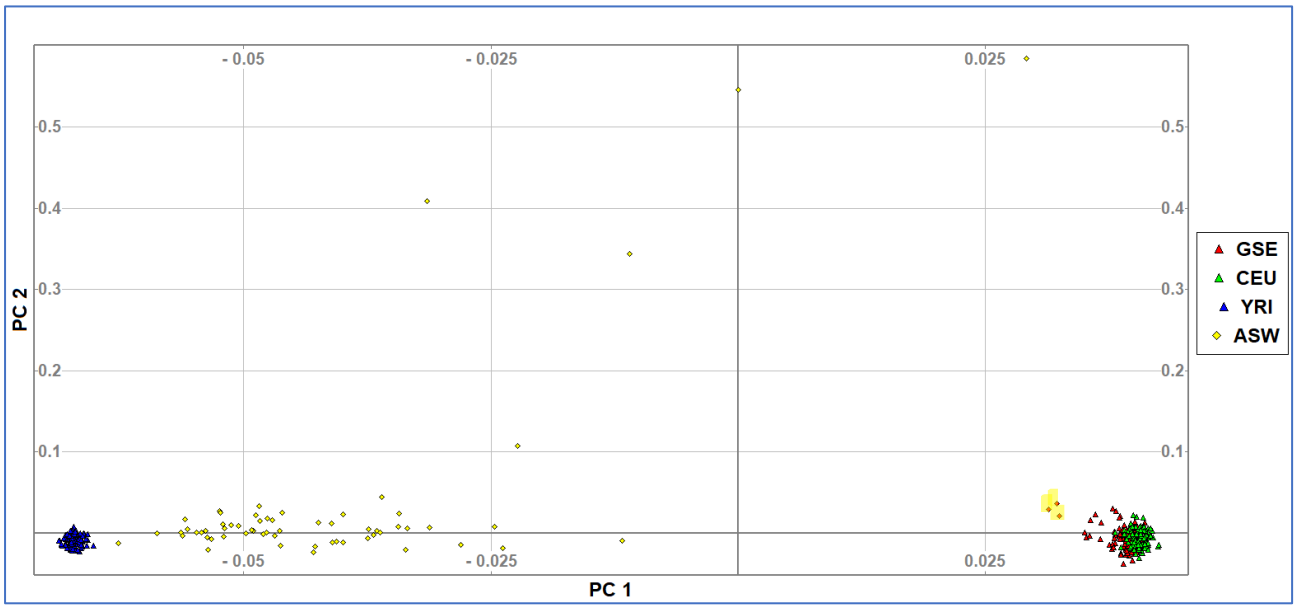


Figure 3.2: PC plot for the dataset (GSE) with CEU (European), ASW (admixed population with some African-ancestry) and YRI (Nigerian) populations from the 1000 Genomes Project. PC plot with PC1 and PC2 for GSE39036 and CEU, ASW and YRI from KGP revealed three outliers (shown in yellow) that were removed from further analysis.

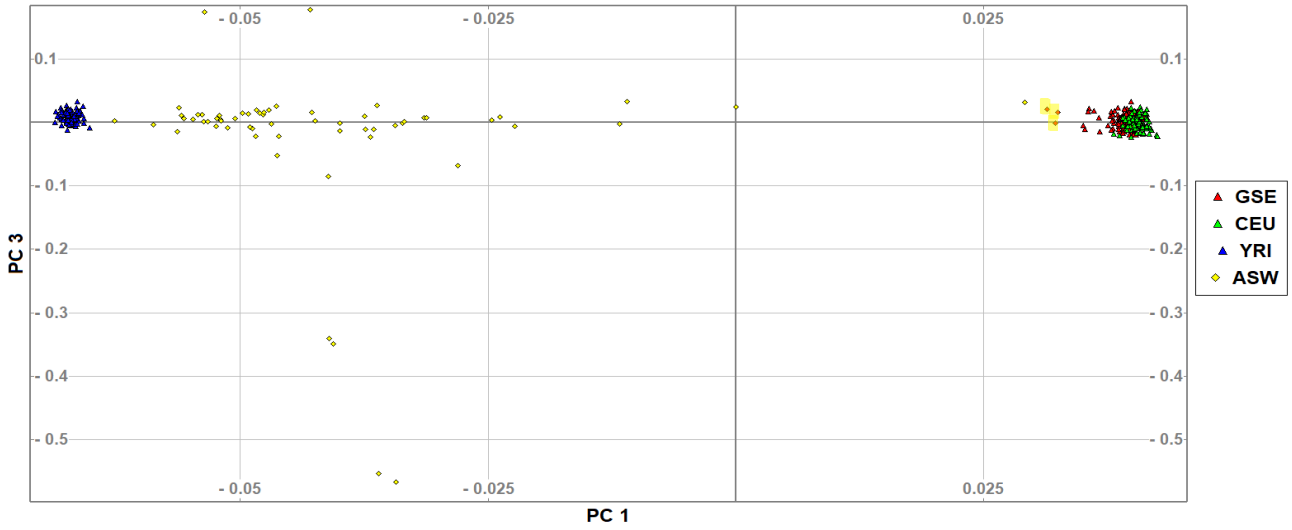


Figure 3.3: PC plot for the dataset (GSE) with CEU (European), ASW (admixed population with some African-ancestry) and YRI (Nigerian) populations from the 1000 Genomes Project. PC plot with PC1 and PC3 for GSE39036 and CEU, ASW and YRI from KGP revealed three outliers (shown in yellow) that were removed from further analysis.

Sample QC of the genotype dataset therefore resulted in the removal of seven samples (GSM954409, GSM954410, GSM954415, GSM954416, GSM954398, GSM954391, GSM954346) from 149 samples. GSM954409, GSM954410, GSM954415 and GSM954416 were removed because of high missingness. GSM954398, GSM954391 and GSM954346 were removed because they are population outliers.

3.2 SNP-RELATED QUALITY CONTROL

In all, 20,652 SNPs from a total 318,237 were excluded from the analysis based on call rates, MAF and departure from HWE. The following were the threshold values used for missingness, minor allele frequency and deviation from HWE:

- --geno 0.065 (SNPs with more than 6.5% missing genotype rate were removed)
- --maf 0.01 (SNPs with a MAF lower than this were removed)
- --hwe 0.00001 (SNPs above this p-value HWE threshold were removed)

Figure 3.5 shows the different steps done in the QC of the genotype data and the number of samples and SNPs removed.

3.3 TRANSCRIPTION DATA

The gene expression dataset, GSE32504, was downloaded from NCBI's GEO and included data from a total of 149 samples. As GSE32504 was normalized, only a sample level QC was done for this dataset. This was achieved by plotting a PC plot in R for the 149 samples. Since they are samples obtained from the liver, they are expected to cluster together.

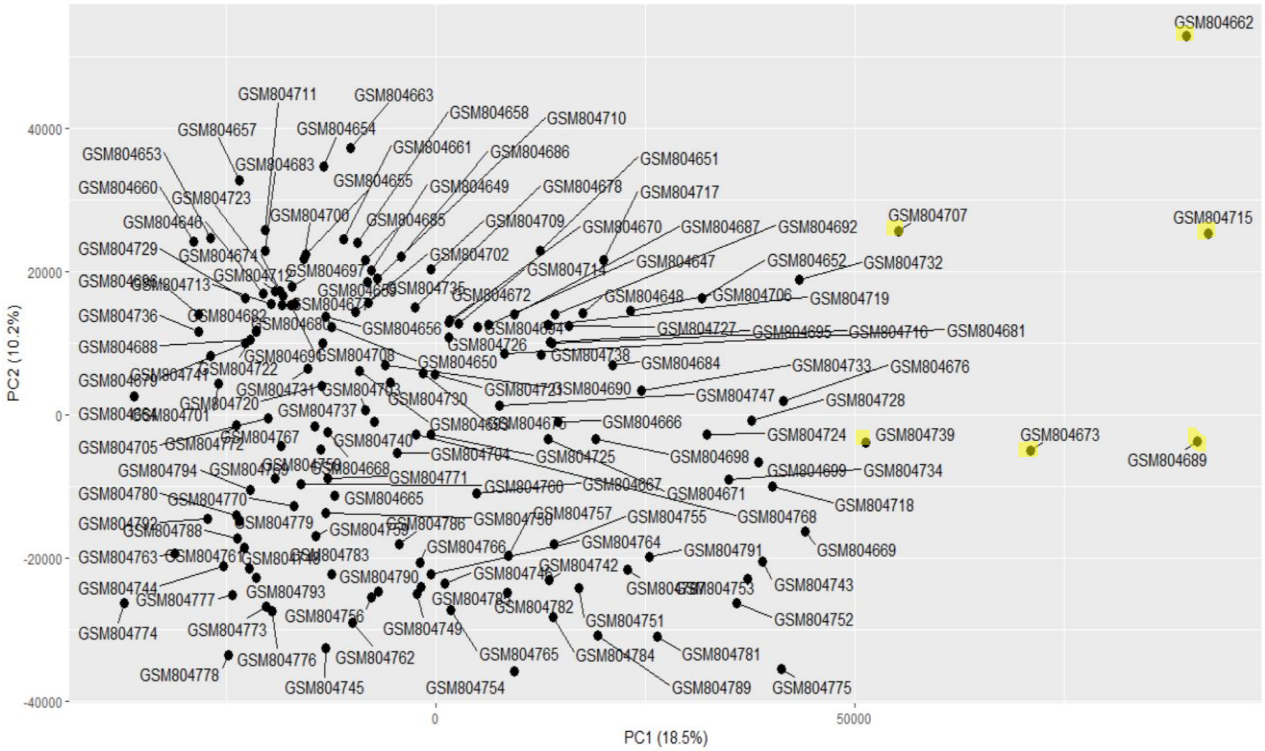


Figure 3.4: Clustering of samples in the gene expression dataset. PC plot for 149 samples revealed six outliers (shown in yellow) that were removed from further analysis.

All samples to the right of 50,000 mark on PC1 were excluded from further analysis. Thus, six samples (GSM804707, GSM804739, GSM804673, GSM804662, GSM804715 and GSM804689) from a total of 149 samples were excluded from further analysis from the gene expression dataset (Figure 3.5).

Since samples in the genotype and gene expression dataset used different sample identifiers, in order to determine which samples failed QC in both datasets, samples from both datasets were mapped using common data variables. After mapping, the samples that failed genotype

QC were identified as 52, 97, 104, 115, 116, 121 and 122. The samples that failed gene expression QC were identified as 17, 28, 44, 62, 70 and 94. Since there was no overlap, a total of 13 samples were excluded from further analysis.

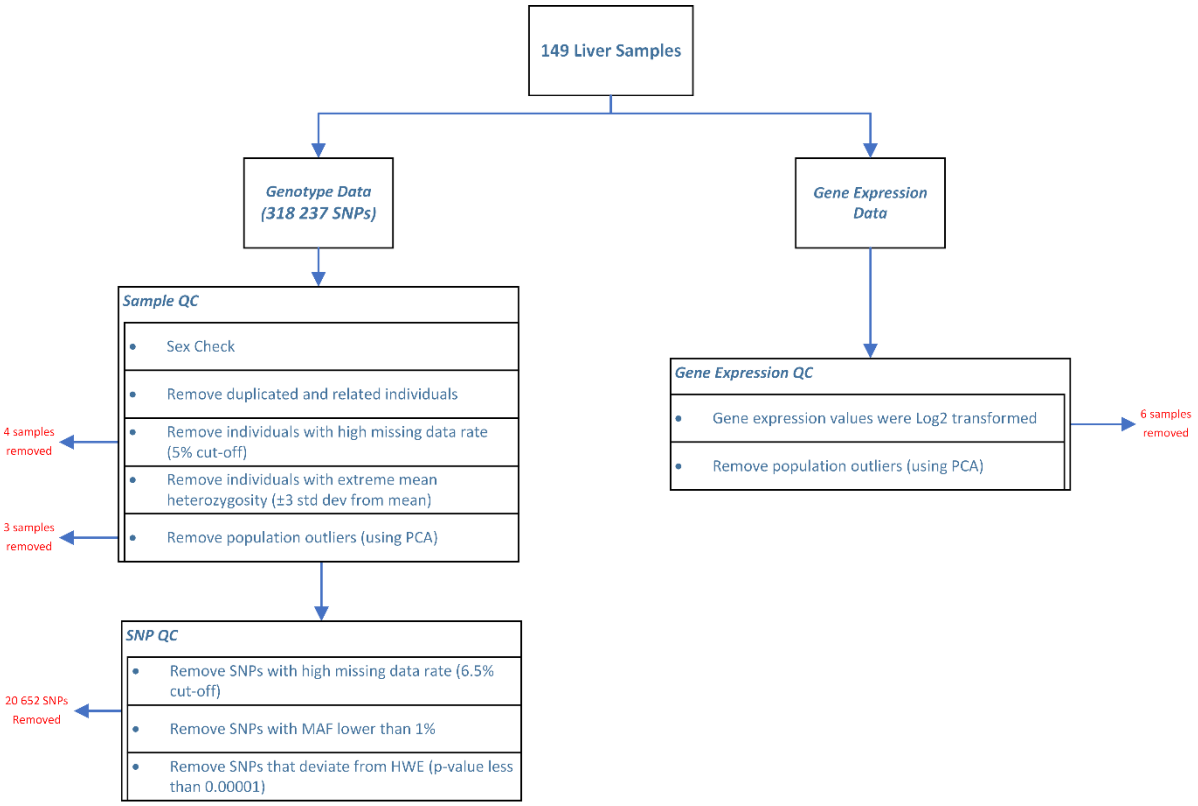


Figure 3.5: Genotype and Gene expression data QC

The bioinformatic pipeline generated to identify eQTLs from genotype and corresponding gene expression datasets can be accessed at https://github.com/phelelani/nf-eqtl. The computing time was assessed once the pipeline had been constructed and tested, and the processes were found to be highly efficient. When the pipeline was run, QC of both the genotype and gene expression datasets took 2m and 36s. Generating the input files needed to run Matrix eQTL took 1m 18s. It took the pipeline 937.91 seconds to generate *cis*- and *trans*-eQTLs from the dataset. From beginning to end, it took the pipeline 1,171.91s (19m and 53s) to generate *cis*- and *trans*-eQTLs.

As imputation of the genotype dataset significantly increases the number of genetic variants included in the study, the pipeline was also run using imputed genotype dataset. Imputation of the genotype dataset was done using the Phase 3 1000 Genomes Project as the reference panel by Michigan imputation server. Imputation of GSE39036 resulted in a total of 9,266,973 SNPs. QC was done post-imputation where SNPs with an imputation score lower than 0.8, SNPs with lower minor allele count bound greater than one and SNPs with greater than 60% missing data were removed. Only SNPs with a minimum and maximum of 2 alleles were tolerated in the QC. Post-QC, 930921 SNPs were removed, and the number of SNPs was reduced to 8,336,052 (Figure 3.6). QC filtering based on HWE and MAF thresholds was not done post-imputation as the dataset is relatively small with 136 samples that passed genotype and transcription data QC. However, a MAF threshold of 1/(136*2), 0.0036, was set when post-imputation, the lower minor allele count bound was set to greater than one. Such a low MAF threshold helps to identify rare eQTLs as well.

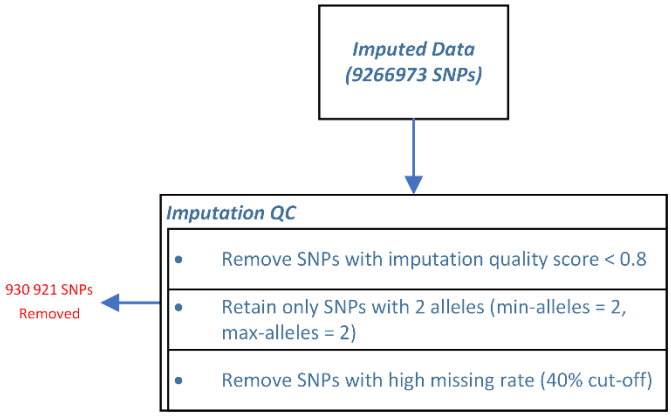


Figure 3.6: Post-imputation QC

As the genotype data was annotated in build hg18 and the imputation dataset in build hg19, in order to compare the eQTLs generated from both datasets, LiftOver was done on the SNP location and the gene location files. 446 genes had to be removed because their gene locations could not be converted to hg19. Similarly, when converting the build of the SNP location file from hg18 to hg19, 31 SNPs had to be removed as their SNP locations could not be converted to hg19. Genotype information of the SNPs that failed LiftOver were also excluded from the genotype file.

To understand what impact imputation and the removal of samples have on the number of eQTLs identified, the number of eQTLs generated after removing either 4 or 13 samples that failed different stages of QC were compared. For that purpose, the dataset and imputed dataset with different number of samples removed will be addressed as the following:

i. Dataset A- Dataset with 4 samples removed (the four samples with high missingness);
ii. Dataset B- Dataset with 13 samples removed (the 13 samples that failed genotype and gene expression QC);
iii. Imputed dataset A- Imputed dataset with 4 samples removed (the same four samples with high missingness removed to generate Dataset A);
iv. Imputed dataset B- Imputed dataset with 13 samples removed (the same 13 samples that were excluded from further analysis to generate Dataset B).

Run 1: The Matrix eQTL package was first run for n=145 samples after removing the four samples with high missingness. The p value threshold for generating eQTLs was set to 1e-6. A FDR less than 0.05 (5%) was chosen to be statistically significant for the linear model that was run with Matrix eQTL. 33 cis-eQTLs and 47,922 trans-eQTLs were generated in 939.45 seconds. The cis-eQTLs generated are shown in Table 3.1. Only 5 cis-eQTLs had an FDR less than 5%.

Run 2: The Matrix eQTL package was run with dataset B (n=136). The p value threshold for identifying eQTLs was set to 1e-6 and FDR had to be less than 0.05 to be statistically significant. The pipeline took 937.91 seconds to run and generated 36 cis-eQTLs and 46,921 trans-eQTLs. The cis-eQTLs are shown in Table 3.2.Only 6 cis-eQTLs had an FDR of less than 5%.

Table 3.1: List of 33 *cis*-eQTLs from Dataset A

SNP	Gene	beta value	t-stat	p-value	FDR
11:5777023	*SMPD1*	305.32	6.36	2.66E-09	0.01
11:5776470	*SMPD1*	-304.05	-6.33	3.12E-09	0.01
18:11213760	*APCDD1*	-30.01	-6.3	3.59E-09	0.01
19:44124368	*KCNN4*	-24.3	-6.18	6.65E-09	0.01
11:66276576	*CD248*	-118.08	-6.09	1.01E-08	0.02
7:99447046	*EPO*	129.73	5.76	5.15E-08	0.06
2:9064191	*DDEF2*	44.14	5.74	5.61E-08	0.06
21:42870522	*TFF3*	-232.73	-5.73	5.85E-08	0.06
12:10196088	*DDX12*	4.54	5.7	6.84E-08	0.06
12:53070392	*ANKRD33*	-790.84	-5.68	7.44E-08	0.06
13:39708950	*LHFP*	65.08	5.62	9.92E-08	0.08
12:104506318	*LOC644278*	3.61	5.58	1.19E-07	0.08
2:9001492	*DDEF2*	-41.86	-5.51	1.62E-07	0.11
2:9100137	*DDEF2*	35.61	5.46	2.08E-07	0.12
19:15254571	*DNAJB1*	-938.05	-5.46	2.12E-07	0.12
12:10188967	*TAS2R10*	-6.83	-5.42	2.51E-07	0.13
8:102109143	*HS.583659*	-5.02	-5.38	2.99E-07	0.15
12:10188967	*DDX12*	-4.49	-5.37	3.18E-07	0.15
22:32783301	*SLC5A1*	-35.34	-5.35	3.44E-07	0.15
7:151024986	*SLC4A2*	-92.52	-5.3	4.36E-07	0.17
18:24370116	*CHST9*	-87.57	-5.29	4.46E-07	0.17
13:39703385	*LHFP*	-52	-5.29	4.61E-07	0.17
7:141736798	*HS.567464*	4.65	5.28	4.85E-07	0.17
5:135469500	*SPOCK*	-21.53	-5.27	4.96E-07	0.17
2:9016356	*DDEF2*	-34.8	-5.26	5.25E-07	0.17
1:89197228	*GBP5*	-125.82	-5.26	5.35E-07	0.17
1:36156432	*SFPQ*	81.58	5.25	5.60E-07	0.18
19:18333273	*UNC13A*	-4.71	-5.24	5.77E-07	0.18
1:36225948	*SFPQ*	79.64	5.19	7.31E-07	0.21
1:213717946	*HS.563744*	12.46	5.17	7.75E-07	0.22
16:54299503	*HS.540546*	-6.62	-5.15	8.44E-07	0.22
10:5425422	*CALML3*	141.34	5.15	8.49E-07	0.22
1:216714314	*USH2A*	-294.33	-5.15	8.53E-07	0.22

Table 3.2: List of 36 *cis*-eQTLs from dataset with Dataset B

SNP	Gene	beta value	t-stat	p-value	FDR
11:5777023	*SMPD1*	309.48	6.47	1.80E-09	0.01
19:44124368	*KCNN4*	-24.77	-6.43	2.13E-09	0.01
11:5776470	*SMPD1*	-308.08	-6.43	2.15E-09	0.01
1:89197228	*GBP5*	-140.04	-6.14	9.10E-09	0.02
12:53070392	*ANKRD33*	-875.02	-5.91	2.77E-08	0.04
13:39708950	*LHFP*	66.09	5.89	3.07E-08	0.04
4:23795064	*HS.581946*	-86.9	-5.62	1.10E-07	0.12
15:52274863	*GNB5*	-33.49	-5.61	1.15E-07	0.12
7:99447046	*EPO*	129.03	5.56	1.43E-07	0.14
13:39703385	*LHFP*	-53.69	-5.5	1.88E-07	0.16
1:186612593	*PDC*	-4.31	-5.39	3.12E-07	0.22
19:15254571	*DNAJB1*	-945.1	-5.36	3.53E-07	0.22
12:104506318	*LOC644278*	3.45	5.36	3.62E-07	0.22
8:141184000	*HS.583577*	4.83	5.35	3.82E-07	0.22
8:102109143	*HS.583659*	-5.25	-5.3	4.64E-07	0.22
18:24370116	*CHST9*	-90.75	-5.3	4.68E-07	0.22
12:10188967	*TAS2R10*	-6.87	-5.3	4.77E-07	0.22
8:74346862	*JPH1*	-57.12	-5.3	4.78E-07	0.22
7:151024986	*SLC4A2*	-94.13	-5.27	5.30E-07	0.22
1:231929075	*LOC440731*	-48.39	-5.27	5.35E-07	0.22
12:10196088	*DDX12*	4.33	5.26	5.74E-07	0.22
19:17672385	*ARRDC2*	-144.23	-5.25	5.88E-07	0.22
9:130175096	*HS.545606*	5.46	5.24	6.27E-07	0.22
16:54299503	*HS.540546*	-6.71	-5.23	6.50E-07	0.22
8:133392693	*KCNQ3*	-9.71	-5.23	6.61E-07	0.22
5:135469500	*SPOCK*	-21.83	-5.22	6.68E-07	0.22
4:36081243	*ARAP2*	-26.03	-5.2	7.50E-07	0.23
8:127524012	*HS.196114*	-3.89	-5.18	8.03E-07	0.23
21:46110544	*TMEM1*	15.71	5.18	8.17E-07	0.23
20:3724708	*PTPRA*	29.4	5.16	8.69E-07	0.23
1:186467274	*PDC*	-4.24	-5.16	8.90E-07	0.23
8:60613335	*CA8*	-25.63	-5.15	9.37E-07	0.23
17:4806052	*DHX33*	-129.02	-5.14	9.76E-07	0.23
3:44489880	*ZNF445*	6.48	5.14	9.80E-07	0.23
2:2583212	*HS.565623*	4.54	5.14	9.85E-07	0.23
7:103015677	*FLJ36166*	-17.03	-5.13	9.90E-07	0.23

Run 3: The Matrix eQTL package was run with imputed dataset A (n= 145, after removing the 4 samples that had high sample missingness). 25,962 *cis*-eQTLs and 38,105,386 *trans*-eQTLs were generated. 16,040 *cis*-eQTLs and 22,867,252 *trans*-eQTLs had FDR less than 5%.

Run 4: The Matrix eQTL package was run with imputed dataset B (n=136, after removing all the 13 samples that failed both genotype and gene expression QC). A total of 25,275 *cis*-eQTLS and 35,837,751 *trans*-eQTLs were generated when 13 samples were excluded from further analysis. Of the 25,275 *cis*-eQTLs, 15,275 *cis*-eQTLs had FDR less than 0.05. 21,743. 649 *trans*-eQTLs had FDR less than 0.05.

3.4 COMBINED SUMMARY ANALYSIS OF THE DIFFERENT RUNS

Dataset A (run 1) generated 33 *cis*-eQTLs and 47922 *trans*-eQTLs. Dataset B (run 2) generated 36 *cis*-eQTLs and 46921 *trans*-eQTLs. There was an overlap of 17 of the *cis*-eQTLs generated by dataset A and B. Table 3.3 provides a summary of the number of *cis*- and *trans*-eQTLs generated from runs 1,2, 3 and 4.

Table 3.3: *Cis*- and *trans*-eQTLs for different dataset inputs

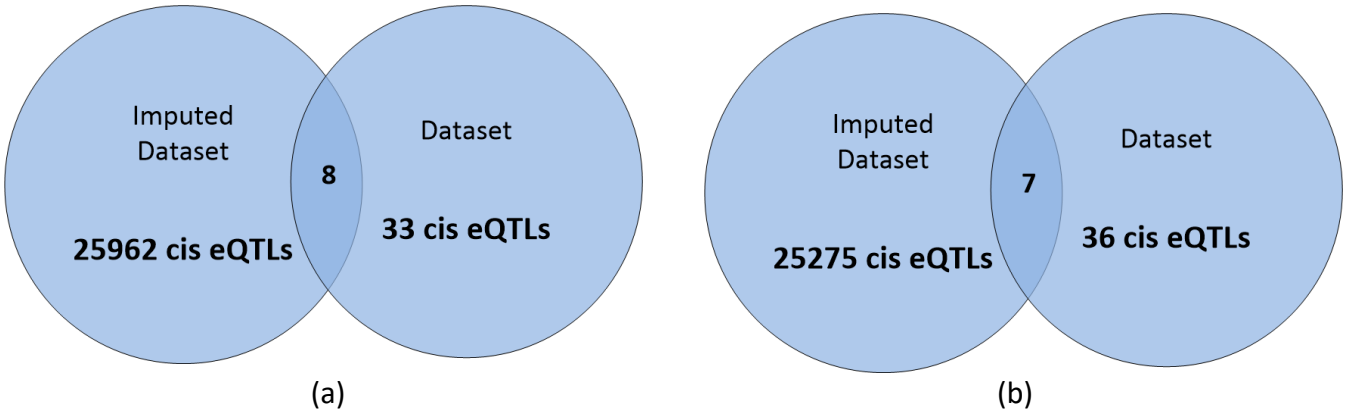
Dataset Used	***Cis* eQTLs**		***Trans* eQTLs**	
	Total No	**FDR <5%**	**Total No**	**FDR <5%**
Dataset A	33	5	47922	8969
Imputed dataset A	25962	16040	38105386	22867252
Dataset B	36	6	46921	6974
Imputed dataset B	25275	15275	35837751	21743649

An assessment was done to determine whether there were any common eQTLs between dataset A and imputed dataset A. From Figure 3.7, we see that only eight *cis*-eQTLs were found to be common between both datasets. Similarly, only seven *cis*-eQTLs were found to be common between dataset B and imputed dataset B. The common *cis*-eQTLs between dataset A and imputed dataset A and between dataset B and imputed dataset B are listed in

Table 3.4. Common *cis*-eQTLs between them both are highlighted in blue. The finding that all *cis*-eQTLs identified in datasets A and B are not present in imputed datasets A and B respectively is somewhat puzzling, as all eQTLs identified in dataset A and B are expected to be present in the post-imputation datasets, imputed dataset A and B respectively. A possible explanation is that the eQTLs not identified in the imputed datasets failed the significance cut off in the larger datasets.

Table 3.4: (A) shows common *cis*-eQTLs between dataset A and imputed dataset A. (B) shows common eQTLs between dataset B and imputed dataset B

(A) Common eQTLs between dataset A and imputed dataset A		**(B) Common eQTLs between dataset B and imputed dataset B**	
eQTL	**Gene**	**eQTL**	**Gene**
11:5777023	*SMPD1*	11:5777023	*SMPD1*
11:5776470	*SMPD1*	11:5776470	*SMPD1*
18:11213760	*APCDD1*	1:231929075	*LOC440731*
19:44124368	*KCNN4*	19:44124368	*KCNN4*
11:66276576	*CD248*	15:52274863	*GNB5*
13:39708950	*LHFP*	13:39708950	*LHFP*
5:135469500	*SPOCK*	5:135469500	*SPOCK*
10:5425422	*CALML3*		



(a) (b)

Figure 3.7: (a)Common *cis*-eQTLs between the dataset A and imputed dataset A. (b) Common *cis*-eQTLs between the dataset B and imputed dataset B

3.5 IDENTIFYING eQTLs FOR CORE AND EXTENDED ADME GENES

The next objective was to identify eQTLs for the core and extended ADME genes from the datasets. For this purpose, the core ADME gene list (32 genes) and the extended ADME gene list (267 genes) from PharmaADME consortium was used. No eQTLs were found for the core ADME genes in datasets A and B. A *cis*-eQTL was identified for an extended ADME gene, from both dataset A and dataset B. However, this eQTL did not meet the FDR criteria of less than 5%. Imputed dataset A generated 6 *cis*-eQTLs for core ADME genes (

Table 3.6), all of which met the FDR criteria. 57 *cis*-eQTLs for extended ADME genes were found in this dataset all of which passed the FDR criteria (Table 3.8). Imputed dataset B produced the same 6 *cis*-eQTLs for core ADME genes (

Table 3.6), all of which met the FDR criteria. 53 *cis*-eQTLs for extended ADME genes were found in this dataset all of which met the FDR criteria (Table 3.9). SNP-nexus was used to identify the rsIDs of the *cis*-eQTLs generated.

Table 3.5 shows the number of *cis*-eQTLs generated for core ADME genes and extended ADME genes in the four different datasets. Figure 3.8 shows that imputed dataset A and imputed dataset B share 49 *cis*-eQTLs for extended ADME genes. Imputed dataset A has 10 unique *cis*-eQTLs and imputed dataset B has 6 unique *cis*-eQTLs.

Table 3.5: *Cis*-eQTLs for core ADME genes and extended ADME genes

Dataset Used	***Cis*-eQTLs**					
	Total No	**FDR <5%**	**Core ADME gene eQTL**	**FDR <5%**	**Extended ADME gene eQTL**	**FDR <5%**
Dataset A	33	5	0	0	1	0
Imputed dataset A	25962	16040	6	6	63	27
Dataset B	36	6	0	0	1	0
Imputed dataset B	25275	15275	6	6	59	28

Table 3.6: *Cis*-eQTLs for core ADME genes from imputed dataset A and B

eQTL	Gene	beta value	t-stat	p-value	FDR	SNP-nexus output
10:97236373	*CYP2C19*	3238.06	5.74	5.64E-08	0.001	rs933765367
7:100085809	*CYP3A5*	16158.9	5.82	3.76E-08	0.001	rs143210517
1:97464825	*DPYD*	325.15	5.16	8.27E-07	0.008	rs114045355
16:27948796	*SULT1A1*	47.61	5.26	5.33E-07	0.005	rs1000598999
2:234515276	*UGT1A1*	1328.46	5.24	5.73E-07	0.006	rs28967009
4:68462245	*UGT2B17*	3944.81	5.54	1.46E-07	0.002	rs917897719

The classification of the core ADME genes, by PharmGKB, for which *cis*-eQTLs were found is shown in Table 3.7. *CYP2C19* and *CP3A5* are Phase I ADME genes. *DPYD* is a Phase I, minor gene. *SULT1A1*, *UGT1A1* and *UGT2B17* are all Phase II ADME genes.

Table 3.7: Classification of the core ADME genes for which eQTLs were found

Gene	PGKB Gene Classification
CYP2C19	Phase I
CYP3A5	Phase I
DPYD	Phase I, minor
SULT1A1	Phase II
UGT1A1	Phase II
UGT2B17	Phase II

Table 3.8: *Cis*-eQTLs for extended ADME genes from imputed dataset A

eQTL	Gene	beta value	t-stat	p-value	FDR	SNP-nexus output
7:150756557	*ABCB8*	97.17	5.56	1.32E-07	0.0017	rs542683997
17:48566159	*ABCC3*	3916.35	5.42	2.54E-07	0.003	rs146347791
13:95531528	*ABCC4*	169.05	7.4	1.09E-11	4.79E-07	rs188869286
16:55140242	*CES1*	5544.24	6.28	4.00E-09	9.30E-05	rs775020102
8:59448296	*CYP7A1*	385.94	7.34	1.51E-11	6.46E-07	rs186995169
1:109812578	*GSTM2*	166.19	5.62	9.66E-08	0.0013	rs200168090

11:64307533	*SLC22A11*	192.52	11.17	3.76E-21	4.92E-16	rs537023014
16:88652760	*SLC7A5*	1438.11	10.32	5.84E-19	6.57E-14	rs78274031
12:21013911	*SLCO1A2*	550.99	5.45	2.17E-07	0.0026	rs151286679
2:235197314	*UGT1A6*	815.66	11.28	2.00E-21	2.67E-16	rs187641587
2:235197314	*UGT1A7*	449.13	8.01	3.81E-13	2.25E-08	rs187641587
2:235197314	*UGT1A8*	84.3	7.59	3.92E-12	1.87E-07	rs187641587
2:235197314	*UGT1A9*	183.07	7.21	3.14E-11	1.26E-06	rs187641587
7:86153073	*ABCB4*	7684.79	5.59	1.11E-07	0.0015	rs555278112
2:220785323	*ABCB6*	870.99	9.75	1.69E-17	1.66E-12	rs143018492
16:47510672	*ABCC11*	109.07	5.7	6.63E-08	0.001	rs74874632
21:43672467	*ABCG1*	-58.98	-7.98	4.67E-13	2.74E-08	rs570922002
15:57764596	*ALDH1A2*	114.95	5.7	6.65E-08	0.001	rs142290839
15:101621567	*ALDH1A3*	356.1	5.94	2.10E-08	0.0004	rs573327784
9:38240923	*ALDH1B1*	863.06	6.6	7.71E-10	2.14E-05	rs192626536
11:66739379	*ALDH3B2*	25.74	5.56	1.31E-07	0.0017	rs148155386
6:134602412	*ALDH8A1*	398.24	6.24	4.85E-09	0.0001	rs571795695
21:36743550	*CBR3*	94	7.4	1.14E-11	4.98E-07	rs551125507
15:75067521	*CYP11A1*	367.74	5.87	2.96E-08	0.0005	rs755222670
2:38109949	*CYP1B1*	320.95	8.1	2.33E-13	1.41E-08	rs568808846
6:31507187	*CYP21A2*	2188.19	10.68	7.05E-20	8.34E-15	rs539689423
20:52156378	*CYP24A1*	39	5.92	2.39E-08	0.0004	rs145004792
10:95133793	*CYP26A1*	283.78	6.46	1.59E-09	4.16E-05	rs138541187
1:171211533	*FMO1*	120.31	7.19	3.49E-11	1.39E-06	rs572768408
1:171015019	*FMO2*	225.67	5.97	1.84E-08	0.0003	rs188999033
1:109276407	*GSTM5*	66.88	5.44	2.34E-07	0.0028	rs558569208
17:37297911	*PNMT*	66.37	7.44	8.83E-12	3.92E-07	rs75035645
6:35514591	*PPARD*	359.5	8.02	3.59E-13	2.13E-08	rs183766594
3:12431128	*PPARG*	830.06	5.52	1.55E-07	0.002	rs571174277
20:44322200	*SLC13A3*	103.53	6.46	1.55E-09	4.05E-05	rs191131632
21:46295837	*SLC19A1*	305.53	6.93	1.36E-10	4.74E-06	rs760591826
11:63884690	*SLC22A12*	56.26	7.23	2.85E-11	1.16E-06	rs140370010
1:115841498	*SLC22A15*	177.01	8.7	8.00E-15	5.64E-10	rs547535334
14:23155228	*SLC22A17*	469.04	8.96	1.74E-15	1.32E-10	rs140088214
15:85328921	*SLC28A1*	132.94	5.55	1.37E-07	0.0018	rs192507939
1:8670986	*SLC2A5*	65.32	5.17	7.71E-07	0.0072	rs564681091
14:22627103	*SLC7A8*	258.18	6.53	1.11E-09	3.00E-05	rs576758648
2:109648855	*SULT1C2*	291.48	6.61	7.22E-10	2.01E-05	rs76376045
6:32844125	*TAP1*	3632.71	7.72	1.91E-12	9.67E-08	rs530350631
2:235197314	*UGT1A10*	421.49	8.94	1.97E-15	1.48E-10	rs187641587
1:47572186	*CYP4Z1*	73.6	5.15	8.42E-07	0.0076	rs1380562224
20:42368032	*HNF4A*	140.87	8.2	1.37E-13	8.68E-09	rs184382741
17:55858396	*MPO*	71.77	7.55	5.00E-12	2.34E-07	rs538785079
7:150369922	*NOS3*	346.06	6.45	1.69E-09	4.40E-05	rs550333765

22:44756769	*SULT4A1*	106.39	11.27	2.15E-21	2.86E-16	rs140325193
6:32844125	*TAP2*	318.86	7.01	8.92E-11	3.32E-06	rs530350631
2:30688726	*XDH*	1542.96	6.02	1.46E-08	0.0003	rs557139987
7:3140989	*CHST12*	187.84	6.06	1.16E-08	0.0002	rs193057824
16:71348081	*CHST4*	943.98	6.32	3.20E-09	1.00E-04	rs146845219
18:24994595	*CHST9*	393.46	5.68	7.59E-08	0.0011	rs117639661
2:169001204	*DHRS9*	571.51	6.77	3.24E-10	1.01E-05	rs192462887
7:118127024	*CFTR*	703.03	8.72	7.00E-15	4.99E-10	rs190428928

Table 3.9: *Cis*-eQTLs for extended ADME genes from imputed dataset B

eQTL	**Gene**	**beta value**	**t-stat**	**p-value**	**FDR**	**SNP-nexus output**
7:150756557	*ABCB8*	97.52	5.52	1.72E-07	0.0022	rs542683997
17:48566159	*ABCC3*	3913.71	5.31	4.44E-07	0.0047	rs146347791
13:95531528	*ABCC4*	170.44	7.55	6.42E-12	2.96E-07	rs188869286
16:55140242	*CES1*	5516.87	6.08	1.23E-08	0.0002	rs775020102
6:52255776	*GSTA3*	40.81	5.42	2.75E-07	0.0032	rs114320112
1:109812578	*GSTM2*	168.06	5.72	6.71E-08	0.0010	rs200168090
11:64307533	*SLC22A11*	192.42	10.85	5.39E-20	6.28E-15	rs537023014
16:88652760	*SLC7A5*	1449.79	10.46	5.20E-19	5.68E-14	rs78274031
12:21013911	*SLCO1A2*	548.69	5.52	1.77E-07	0.0022	rs151286679
2:235197314	*UGT1A6*	815.59	11.02	2.02E-20	2.42E-15	rs187641587
2:235197314	*UGT1A7*	447.83	7.86	1.21E-12	6.57E-08	rs187641587
2:235197314	*UGT1A8*	84.39	7.47	9.54E-12	4.14E-07	rs187641587
2:235197314	*UGT1A9*	182.72	7.05	9.11E-11	3.19E-06	rs187641587
7:86153073	*ABCB4*	7688.9	5.53	1.64E-07	0.0021	rs555278112
2:220785323	*ABCB6*	871.64	9.5	1.25E-16	1.15E-11	rs143018492
16:47510672	*ABCC11*	109.11	5.76	5.63E-08	0.0009	rs74874632
21:43672467	*ABCG1*	-59.15	-7.93	7.96E-13	4.46E-08	rs570922002
15:101621567	*ALDH1A3*	358.08	6.17	7.66E-09	0.0002	rs573327784
9:38240923	*ALDH1B1*	866.58	6.5	1.51E-09	3.86E-05	rs192626536
11:66739379	*ALDH3B2*	26.17	5.67	8.67E-08	0.0013	rs148155386
6:134602412	*ALDH8A1*	399.67	6.1	1.10E-08	0.0002	rs571795695
21:36743550	*CBR3*	94.54	7.45	1.11E-11	4.75E-07	rs551125507
15:75067521	*CYP11A1*	369.12	6.06	1.35E-08	0.0003	rs755222670
2:38109949	*CYP1B1*	322.78	8.18	2.07E-13	1.30E-08	rs568808846
6:31507187	*CYP21A2*	2187.61	10.41	6.62E-19	7.16E-14	rs539689423
20:52156378	*CYP24A1*	39.57	6.16	8.25E-09	0.0002	rs145004792
10:94503207	*CYP26A1*	135.87	6.71	5.20E-10	1.54E-05	rs11817295
1:171211533	*FMO1*	120.57	7.05	8.92E-11	3.13E-06	rs572768408
1:171015019	*FMO2*	224.16	5.81	4.52E-08	0.0007	rs188999033
1:109276407	*GSTM5*	66.81	5.44	2.55E-07	0.0030	rs558569208
17:37297911	*PNMT*	66.41	7.32	2.18E-11	8.79E-07	rs75035645

6:35514591	*PPARD*	359.13	8.31	1.01E-13	6.63E-09	rs183766594
3:12431128	*PPARG*	837.46	5.55	1.50E-07	0.0019	rs571174277
20:44322200	*SLC13A3*	104.2	6.54	1.22E-09	3.19E-05	rs191131632
21:46295837	*SLC19A1*	306.39	6.8	3.21E-10	9.87E-06	rs760591826
11:63884690	*SLC22A12*	56.76	7.27	2.87E-11	1.14E-06	rs140370010
1:115841498	*SLC22A15*	178.1	8.69	1.21E-14	8.63E-10	rs547535334
14:23155228	*SLC22A17*	468.43	8.9	3.80E-15	2.85E-10	rs140088214
15:85328921	*SLC28A1*	133.66	5.54	1.56E-07	0.0020	rs192507939
1:8108538	*SLC2A5*	60.26	5.2	7.39E-07	0.0071	rs181234655
14:22627103	*SLC7A8*	261.67	6.9	1.96E-10	6.27E-06	rs576758648
2:109648855	*SULT1C2*	291.71	6.43	2.15E-09	5.25E-05	rs76376045
2:235197314	*UGT1A10*	420.81	8.73	9.61E-15	6.93E-10	rs187641587
20:42368032	*HNF4A*	141.37	8.25	1.38E-13	8.75E-09	rs184382741
17:55858396	*MPO*	72.75	7.89	1.01E-12	5.55E-08	rs538785079
7:150369922	*NOS3*	343.37	6.52	1.34E-09	3.46E-05	rs550333765
6:159103199	*SOD2*	73.51	6.27	4.79E-09	0.0001	rs139101291
22:44756769	*SULT4A1*	106.99	11.33	3.28E-21	4.28E-16	rs140325193
2:30688726	*XDH*	1539.5	5.92	2.61E-08	0.0005	rs557139987
7:3140989	*CHST12*	191.13	6.47	1.79E-09	4.45E-05	rs193057824
16:71348081	*CHST4*	950.3	6.51	1.44E-09	3.69E-05	rs146845219
18:24571094	*CHST9*	400.48	5.76	5.68E-08	0.0009	rs578033838
2:169058123	*DHRS9*	415.11	5.33	4.11E-07	0.0045	rs145911692

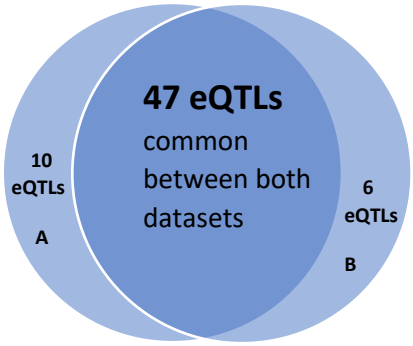


Figure 3.8: Common *cis*-eQTLs for extended ADME genes between imputed dataset A and imputed dataset B. 92.5 % of *cis*-eQTLs identified in imputed dataset B is shared with imputed dataset A.

3.6 EXPRESSION LEVELS OF CORE AND EXTENDED ADME GENES IN THE LIVER

The core and extended ADME genes for which eQTLs were found in this study were searched for in the GTEx database to confirm that they had previously been shown to the be expressed in the liver. Table 3.10 lists the expression levels of these 65 genes in the liver (n= 226 livers). Expression values are shown in TPM. It is calculated from a gene model with isoforms collapsed to a single gene. No other normalisation steps were applied to the expression values. *CES1* (TPM=403.2) is the most highly expressed gene in the liver followed by *CYP3A5* (TPM=155.9) and *ALDH8A1* (TPM=91.16). Though *CYP24A1*, *GSTA3*, *UGT1A8* and *UGT1A10* are assumed to be expressed in the liver, they were found to have no significant expression levels in the 226 liver samples available in GTEx. There was no data available on the expression levels of *CYP11A1*, *GSTM5* and *SLC22A17* in GTEx.

Table 3.10: Expression levels of the core and extended ADME genes in the liver

Gene	Expressed in Liver	TPM (Linear scale)
ABCB4	Yes	28.66
ABCB8	Yes	9.999
ABCC3	Yes	68.68
ABCC11	Yes	1.335
ALDH1A3	Yes	0.6607
ALDH3B2	Yes	0.03769
ALDH8A1	Yes	91.16
CES1	Yes	403.2
CHST9	Yes	0.2595
CYP11A1	No data available	
CYP24A1	No	0
DHRS9	Yes	0.1997
FMO2	Yes	1.736
GSTA3	No	0
GSTM2	Yes	2.488
GSTM5	No data available	
PPARG	Yes	3.55
SLC2A5	Yes	0.1254
SLC28A1	Yes	51.73
SLCO1A2	Yes	0.6809
SOD2	Yes	85.32

XDH	Yes	26.05
CYP2C19	Yes	13.25
CYP3A5	Yes	155.9
DPYD	Yes	12.42
SULT1A1	Yes	54.88
UGT1A1	Yes	22.55
UGT2B17	Yes	4.198
UGT1A6	Yes	2.438
UGT1A7	Yes	0.06162
UGT1A8	No	0
UGT1A9	Yes	10.16
CYP26A1	Yes	4.505
FMO1	Yes	0.1465
SLC13A3	Yes	5.066
SLC19A1	Yes	6.343
SLC22A12	Yes	0.02704
SLC22A15	Yes	0.3342
SLC22A17	No data available	
SLC7A8	Yes	1.349
SULT1C2	Yes	0.386
UGT1A10	No	0
HNF4A	Yes	55.41
MPO	Yes	0.4426
NOS3	Yes	1.829
SULT4A1	Yes	0.1532
CYP7A1	Yes	2.612
ALDH1A2	Yes	1.377
TAP1	Yes	13.19
CYP4Z1	Yes	0.3359
TAP2	Yes	6.756
CFTR	Yes	0.5618
CHST12	Yes	1.003
CHST4	Yes	1.702
PNMT	Yes	0.1801
PPARD	Yes	12.66
CYP1B1	Yes	2.596
CYP21A2	Yes	17.46
ABCB6	Yes	24.8
CBR3	Yes	0.4145
ABCG1	Yes	1.619
ALDH1B1	Yes	40.11
SLC22A11	Yes	0.1209
SLC7A5	Yes	13.75
ABCC4	Yes	0.4402

3.7 VALIDATION OF eQTLs IDENTIFIED USING GTEx AND QTLbase DATABASE

The eQTLs identified for the core and extended ADME gene were not found in the GTEx database. When QTLbase database was queried for the core ADME genes under the 'Traits' tab, *cis*-eQTLs were identified for all the six core ADME genes in the liver (1 eQTL for *CYP2C19*, 611 for *CYP3A5*, 31 for *DPYD*, 6 for *SULT1A1*, 100 for *UGT1A1* and 15 for *UGT2B17*). However, the eQTLs identified by this study were not replicated for any of the six core ADME genes.

3.8 VISUALISATION OF GENE EXPRESSION DISTRIBUTION BY eQTLs

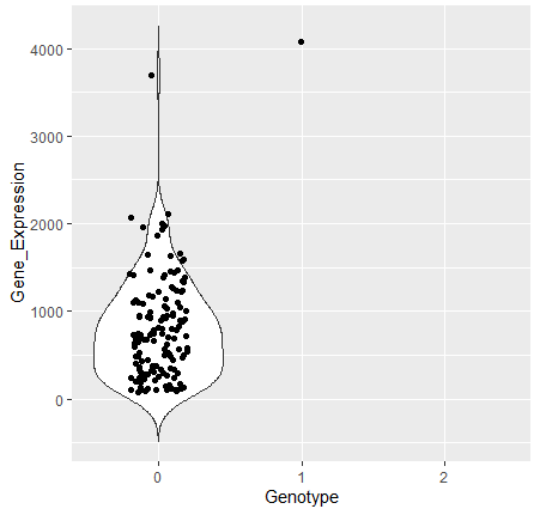
The genotype distribution of the eQTLs for the six core ADME genes is shown in Table 3.11. Genotype 0 corresponds to homozygous major (AA), genotype 1 corresponds to heterozygous (AB) and genotype 2 corresponds to homozygous minor (BB).

Figure 3.9 shows violin plots for *cis*-eQTLs and the ADME gene it affects. They show how genotype of the eQTL affects the expression of the genes in question.

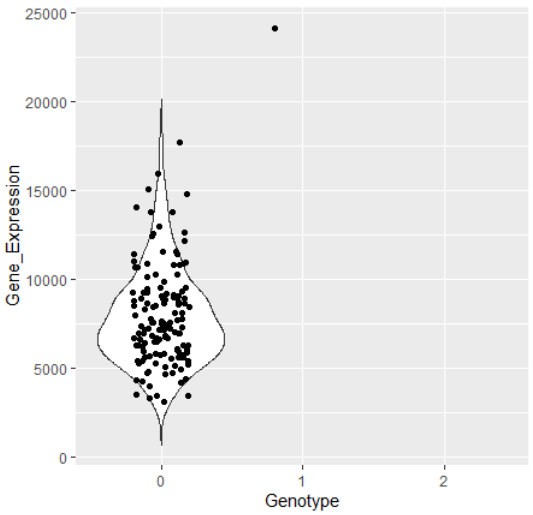
Table 3.11: Genotype distribution of eQTLs for core ADME genes

eQTL	**rsID**	**Core ADME gene**	**Genotype: 0**	**Genotype: 1**	**Genotype: 2**
10:97236373	rs933765367	*CYP2C19*	144	1	NA
7:100085809	rs143210517	*CYP3A5*	144	1	NA
1:97464825	rs114045355	*DPYD*	140	5	NA
16:27948796	rs1000598999	*SULT1A1*	144	1	NA
2:30688726	rs28967009	*UGT1A1*	144	1	NA
4:68462245	rs917897719	*UGT2B17*	144	1	NA

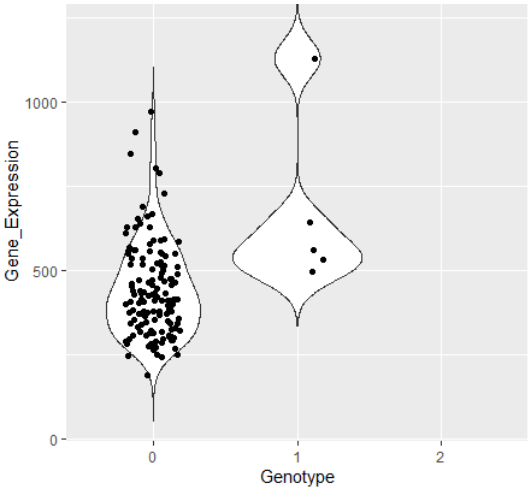
It can be seen in Figure 3.9 that for all the core ADME genes except for *DPYD*, most of the samples show the homozygous major genotype for the eQTL in question. The variant allele frequencies for the eQTLs were very low thereby affecting the power to confirm them as robust eQTLs.



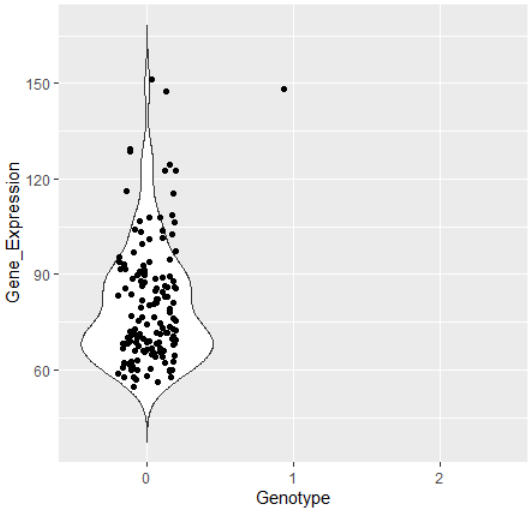
a) rs933765367 and *CYP2C19*



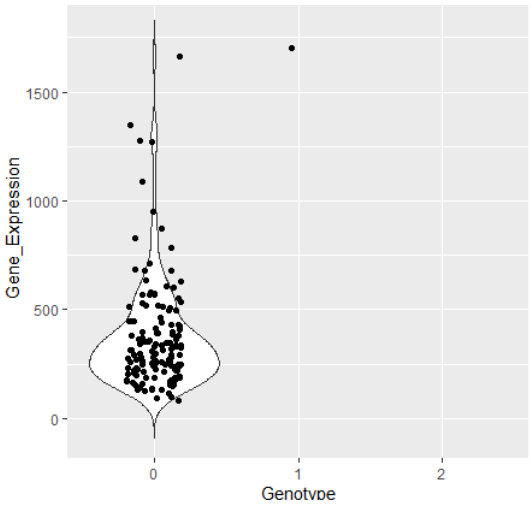
b) rs143210517 and *CYP3A5*



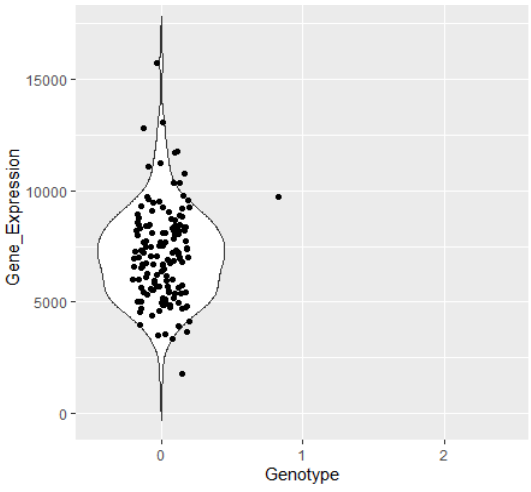
c) rs114045355 and *DPYD*



d) rs1000598999 and *SULT1A1*



e) rs28967009 and *UGT1A1*



f) rs917897719 and *UGT2B17*

Figure 3.9: Violin plots generated to show how eQTL genotype affects the expression of the corresponding ADME gene. a) depicts how rs933765367 genotype affects *CYP2C19* expression. b) depicts how rs143210517 genotype affects *CYP3A5* expression. c) depicts how rs114045355 genotype affects *DPYD* expression. d) depicts how rs1000598999 genotype

affects *SULT1A1* expression. e) depicts how rs28967009 genotype affects *UGT1A1* expression. f) depicts how rs917897719 genotype affects *UGT2B17* expression.

The genotype distribution of the eQTLs for the sixty-three extended ADME genes is shown in Table 3.12. As described earlier, genotype 0 corresponds to homozygous major (AA), 1 corresponds to heterozygous (AB) and 2 corresponds to homozygous minor (BB).

Table 3.12: Genotype distribution of eQTLs for extended ADME genes

eQTL	**rsID**	**Extended ADME gene**	**Genotype:0**	**Genotype:1**	**Genotype:2**
7:150756557	rs542683997	*ABCB8*	144	1	NA
17:48566159	rs146347791	*ABCC3*	143	2	NA
13:95531528	rs188869286	*ABCC4*	144	1	NA
16:55140242	rs775020102	*CES1*	144	1	NA
8:59448296	rs186995169	*CYP7A1*	144	1	NA
1:109812578	rs200168090	*GSTM2*	144	1	NA
11:64307533	rs537023014	*SLC22A11*	144	1	NA
16:88652760	rs78274031	*SLC7A5*	144	1	NA
12:21013911	rs151286679	*SLCO1A2*	144	1	NA
2:235197314	rs187641587	*UGT1A6*	144	1	NA
2:235197314	rs187641587	*UGT1A7*	144	1	NA
2:235197314	rs187641587	*UGT1A8*	144	1	NA
2:235197314	rs187641587	*UGT1A9*	144	1	NA
7:86153073	rs555278112	*ABCB4*	144	1	NA
2:220785323	rs143018492	*ABCB6*	144	1	NA
16:47510672	rs74874632	*ABCC11*	144	1	NA
21:43672467	rs570922002	*ABCG1*	144	1	NA
15:57764596	rs142290839	*ALDH1A2*	142	3	NA
15:101621567	rs573327784	*ALDH1A3*	144	1	NA
9:38240923	rs192626536	*ALDH1B1*	144	1	NA
11:66739379	rs148155386	*ALDH3B2*	143	2	NA
6:134602412	rs571795695	*ALDH8A1*	143	2	NA
21:36743550	rs551125507	*CBR3*	144	1	NA
15:75067521	rs755222670	*CYP11A1*	144	1	NA
2:38109949	rs568808846	*CYP1B1*	143	2	NA
6:31507187	rs539689423	*CYP21A2*	144	1	NA
20:52156378	rs145004792	*CYP24A1*	144	1	NA
10:95133793	rs138541187	*CYP26A1*	144	1	NA
1:171211533	rs572768408	*FMO1*	144	1	NA
1:171015019	rs188999033	*FMO2*	143	2	NA
1:109276407	rs558569208	*GSTM5*	144	1	NA

eQTL	rsID	Extended ADME gene	Genotype:0	Genotype:1	Genotype:2
17:37297911	rs75035645	*PNMT*	144	1	NA
6:35514591	rs183766594	*PPARD*	144	1	NA
3:12431128	rs571174277	*PPARG*	144	1	NA
20:44322200	rs191131632	*SLC13A3*	144	1	NA
21:46295837	rs760591826	*SLC19A1*	144	1	NA
11:63884690	rs140370010	*SLC22A12*	144	1	NA
1:115841498	rs547535334	*SLC22A15*	144	1	NA
14:23155228	rs140088214	*SLC22A17*	144	1	NA
15:85328921	rs192507939	*SLC28A1*	144	1	NA
1:8670986	rs564681091	*SLC2A5*	144	1	NA
14:22627103	rs576758648	*SLC7A8*	144	1	NA
2:109648855	rs76376045	*SULT1C2*	144	1	NA
6:32844125	rs530350631	*TAP1*	144	1	NA
2:235197314	rs187641587	*UGT1A10*	144	1	NA
1:47572186	rs1380562224	*CYP4Z1*	144	1	NA
20:42368032	rs184382741	*HNF4A*	144	1	NA
17:55858396	rs538785079	*MPO*	143	2	NA
7:150369922	rs550333765	*NOS3*	144	1	NA
22:44756769	rs140325193	*SULT4A1*	144	1	NA
6:32844125	rs530350631	*TAP2*	144	1	NA
2:30688726	rs557139987	*XDH*	144	1	NA
7:3140989	rs193057824	*CHST12*	143	2	NA
16:71348081	rs146845219	*CHST4*	143	2	NA
18:24994595	rs117639661	*CHST9*	144	1	NA
2:169001204	rs192462887	*DHRS9*	144	1	NA
7:118127024	rs190428928	*CFTR*	144	1	NA
6:52255776	rs114320112	*GSTA3*	138	7	NA
10:94503207	rs11817295	*CYP26A1*	139	6	NA
1:8108538	rs181234655	*SLC2A5*	144	1	NA
6:159103199	rs139101291	*SOD2*	141	4	NA
18:24571094	rs578033838	*CHST9*	143	2	NA
2:169058123	rs145911692	*DHRS9*	144	1	NA

The Appendix shows violin plots for all *cis*-eQTLs identified for extended ADME genes. Once again, the variant alleles of the identified *cis*-eQTLs occurred at very low frequencies, limiting our power to assess them as robust eQTLs.

3.9 ALLELE FREQUENCIES OF *cis*-eQTLs IN AFRICAN POPULATIONS

Allele frequencies of *cis*-eQTLs for the core and extended ADME genes were obtained from African populations from the 1000 Genomes Project and the African Genome Variation Project (n=320). The different African populations and population codes used are listed in Table 3.13. The *cis*-eQTL for *UGT1A1*, rs28967009 (2:234515276) was detected in the GWD, MSL, ESN, YRI, LWK, ASW, ACB, ZUL, BAG and ETH (Table 3.14, Figure 3.10). The *cis*-eQTL for *DPYD*, rs114045355 (1:97464825) was only detected in ASW, an admixed African-ancestry population (Table 3.15).

Table 3.13: Population codes explained

Population code	**Population**
GWD	Gambian in Western Divisions in the Gambia
MSL	Mende in Sierra Leone
ESN	Esan in Nigeria
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
ASW	American's of African Ancestry in SW USA (admixed)
ACB	African Caribbean in Barbados (admixed)
ZUL	Zulu
BAG	Baganda
ETH	Ethiopia

Table 3.14: Allele frequencies of rs28967009 (2:234515276), eQTL for *UGT1A1* in different African populations

eQTL for *UGT1A1*: rs28967009 (2:234515276)			
Dataset	Population	MAF	Sample size
KGP	GWD	0.031	113
	MSL	0.018	85
	ESN	0.061	99
	YRI	0.060	108
	LWK	0.081	99
	ASW	0.033	61
	ACB	0.042	96
AGVP	ZUL	0.080	100
	BAG	0.125	100
	ETH	0.025	120

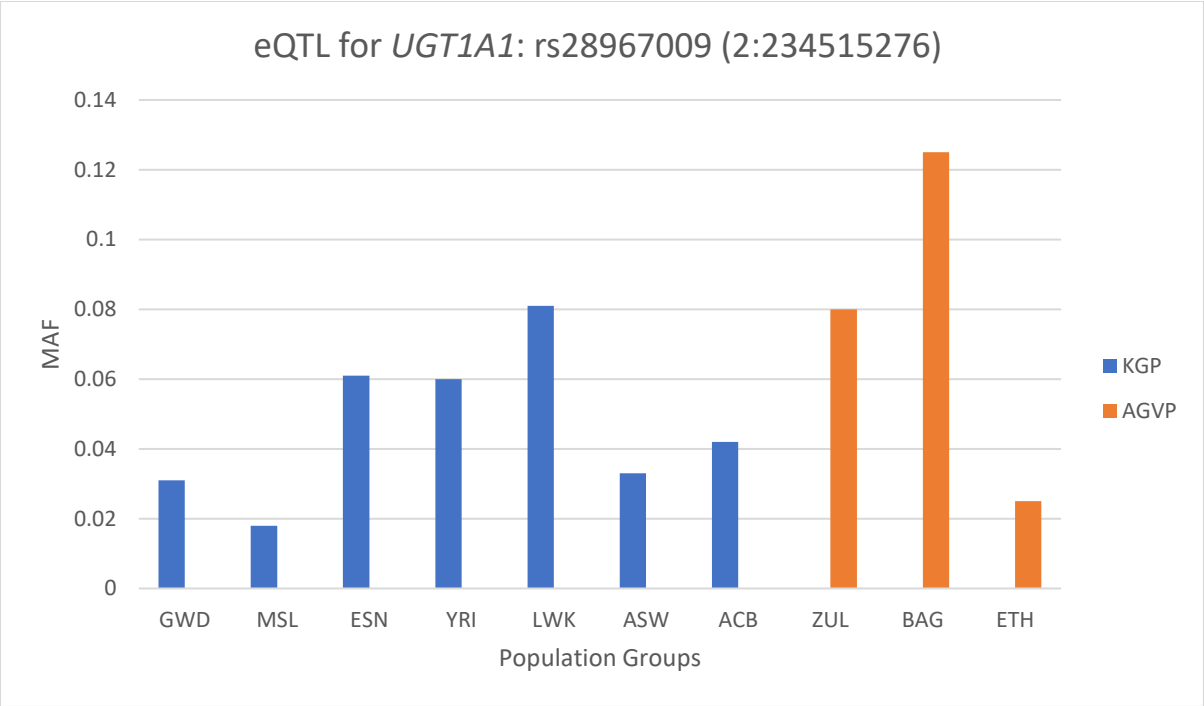


Figure 3.10: Allele frequencies of rs28967009 (2:234515276), eQTL for *UGT1A1* in different African populations

Table 3.15: Allele frequency of rs114045355 (1:97464825), eQTL for *DPYD* in African populations

eQTL for *DPYD*: rs114045355 (1:97464825)			
Dataset	Population	MAF	Sample size
KGP	ASW	0.008	96

Allele frequencies of *cis*-eQTLs for the extended ADME genes were also searched in the different African populations listed in Table 3.13. The following *cis*-eQTLs rs139101291 (eQTL for *SOD2,* Table 3.16, Figure 3.11) , rs11817295 (eQTL for *CYP26A1,*Table 3.17, Figure 3.12), rs191131632 (eQTL for *SLC13A3*, Table 3.18, Figure 3.13), rs146347791 (eQTL for *ABCC3*, Table 3.19, Figure 3.14), rs140088214 (eQTL for *SLC22A17*, Table 3.20, Figure 3.15), rs78274031 (eQTL for *SLC7A5*, Table 3.21, Figure 3.16), rs142290839 (eQTL for *ALDH1A2*, Table 3.22, Figure 3.17), rs117639661 (eQTL for *CHST9*, Table 3.23), rs151286679 (eQTL for *SLCO1A2,* Table 3.24, Figure 3.18), rs74874632 (eQTL for *ABCC11,* Table 3.25), rs542683997

(eQTL for *ABCB8*, Table 3.26) and rs551125507 (eQTL for *CBR3*, Table 3.27) were found to be present in various African populations.

Table 3.16: Allele frequencies of rs139101291 (6:159103199), eQTL for *SOD2* in African populations

eQTL for *SOD2*: rs139101291 (6:159103199)			
Dataset	Population	MAF	Sample size
KGP	GWD	0.32	113
	MSL	0.27	85
	ESN	0.25	99
	YRI	0.32	108
	LWK	0.27	99
	ASW	0.25	61
	ACB	0.27	96
AGVP	ZUL	0.25	100
	BAG	0.31	100
	ETH	0.28	120

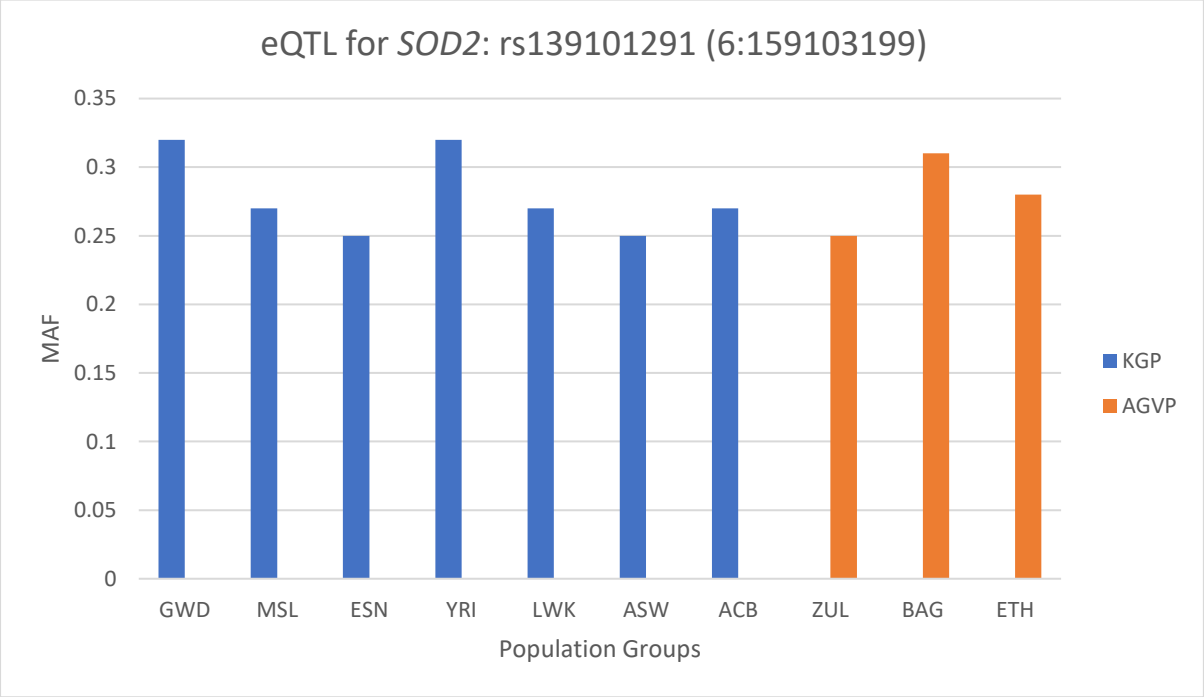


Figure 3.11: Allele frequencies of rs139101291 (6:159103199), eQTL for *SOD2* in African populations

Table 3.17: Allele frequencies of rs11817295 (10:94503207), eQTL for *CYP26A1* in African populations

eQTL for *CYP26A1*: rs11817295 (10:94503207)			
Dataset	Population	MAF	Sample size
KGP	GWD	0.21	113
	MSL	0.13	85
	ESN	0.23	99
	YRI	0.23	108
	LWK	0.20	99
	ASW	0.17	61
	ACB	0.22	96
AGVP	ZUL	0.21	100
	BAG	0.17	100
	ETH	0.23	120

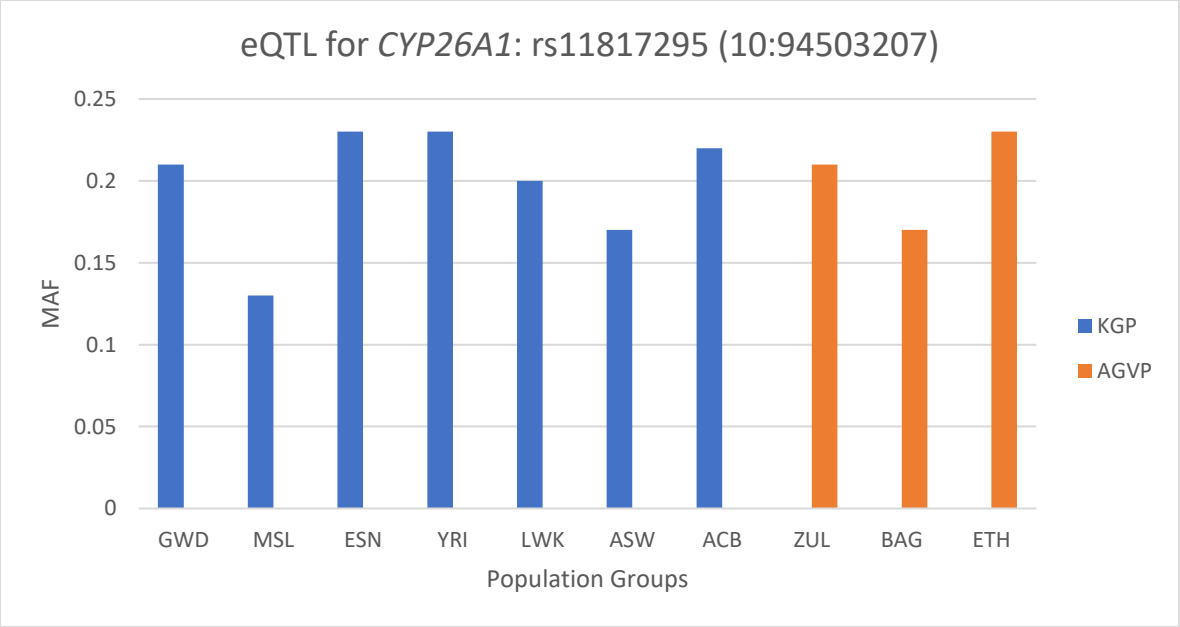


Figure 3.12: Allele frequencies of rs11817295 (10:94503207), eQTL for *CYP26A1* in African populations

Table 3.18: Allele frequencies of rs191131632 (20:44322200), eQTL for *SLC13A3* in African populations

eQTL for *SLC13A3*: rs191131632 (20:44322200)			
Dataset	Population	MAF	Sample size
KGP	GWD	0.04	113
	MSL	0.06	85
	ESN	0.05	99
	YRI	0.02	108
	LWK	0.05	99
	ASW	0.02	61
	ACB	0.06	96
AGVP	ZUL	0.03	100
	BAG	0.07	100
	ETH	0.03	120

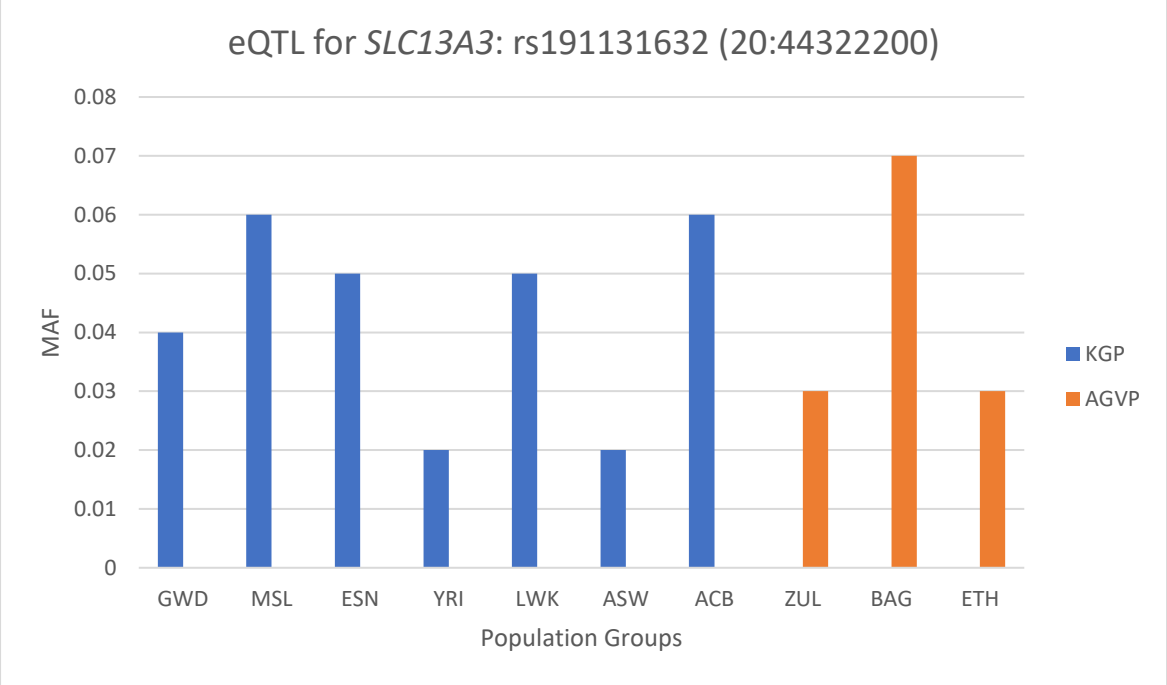


Figure 3.13: Allele frequencies of rs191131632 (20:44322200), eQTL for *SLC13A3* in African populations

Table 3.19: Allele frequencies of rs146347791 (17:48566159), eQTL for *ABCC3* in African populations

eQTL for *ABCC3*: rs146347791 (17:48566159)			
Dataset	Population	MAF	Sample size
KGP	ASW	0.02	61
	LWK	0.03	99
	YRI	0	108
AGVP	ZUL	0.04	100
	BAG	0.01	100

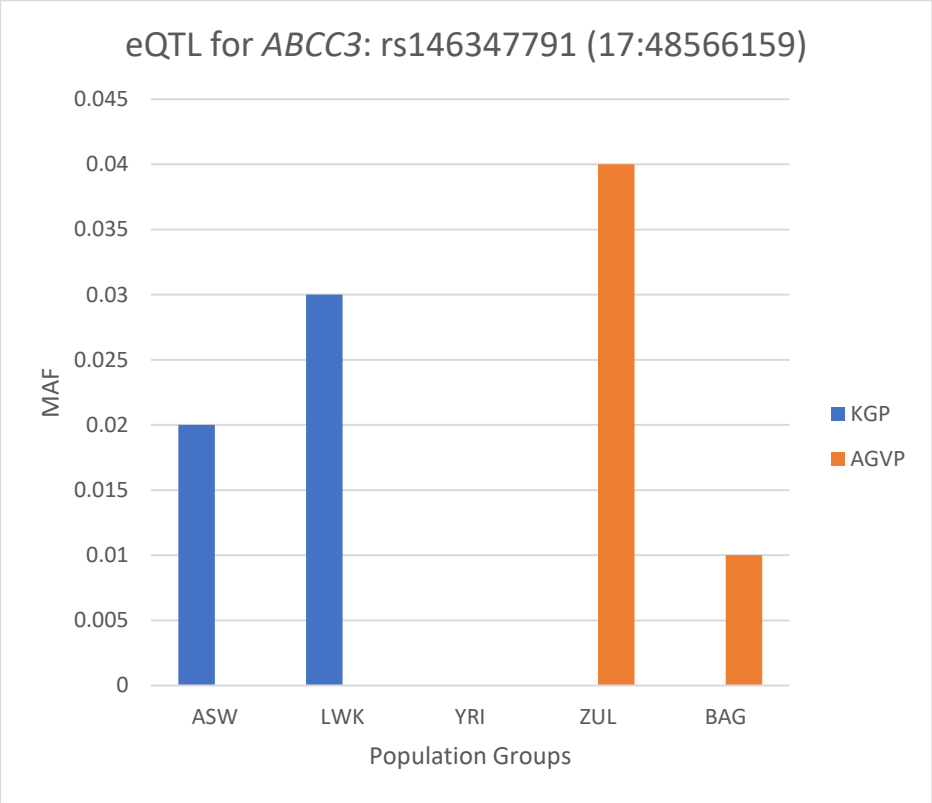


Figure 3.14: Allele frequencies of rs146347791 (17:48566159), eQTL for *ABCC3* in African populations

Table 3.20: Allele frequencies of rs140088214 (14:23155228), eQTL for *SLC22A17* in African populations

eQTL for *SLC22A17*: rs140088214 (14:23155228)			
Dataset	Population	MAF	Sample size
KGP	ASW	0.01	61
	ACB	0.01	96
	LWK	0.01	99
AGVP	ETH	0.03	120

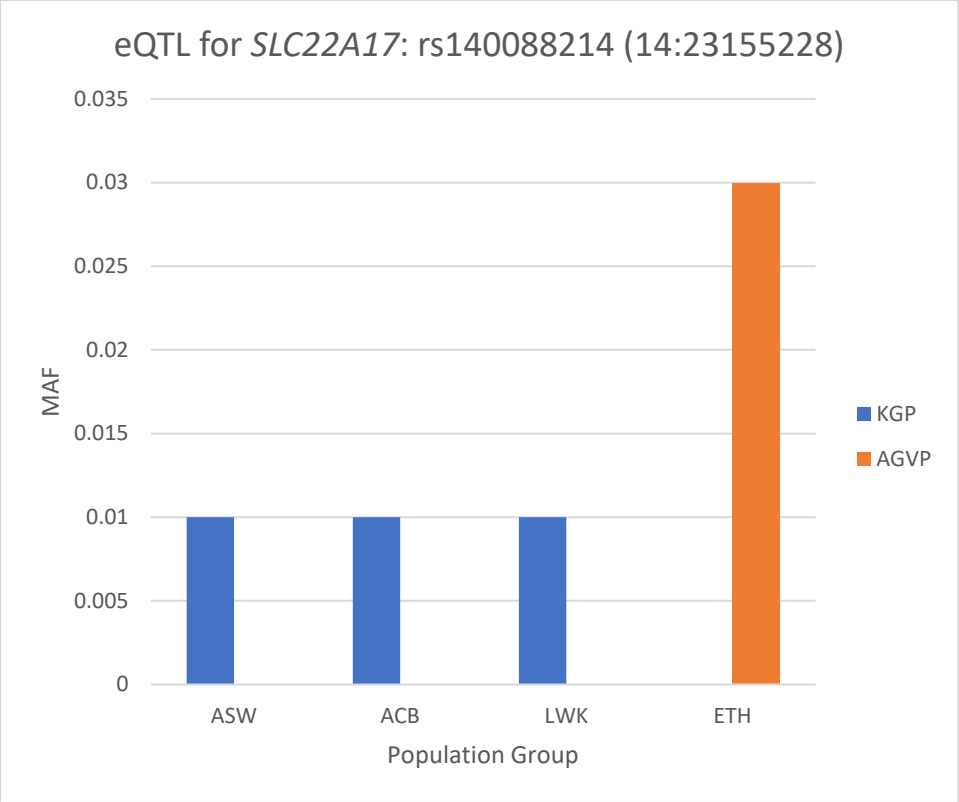


Figure 3.15: Allele frequencies of rs140088214 (14:23155228), eQTL for *SLC22A17* in African populations

Table 3.21: Allele frequencies of rs78274031 (16:88652760), eQTL for *SLC7A5* in African populations

eQTL for *SLC7A5*: rs78274031 (16:88652760)			
Dataset	Population	MAF	Sample size
KGP	ACB	0.01	96
AGVP	ZUL	0.03	100
	BAG	0.01	100
	ETH	0.03	120

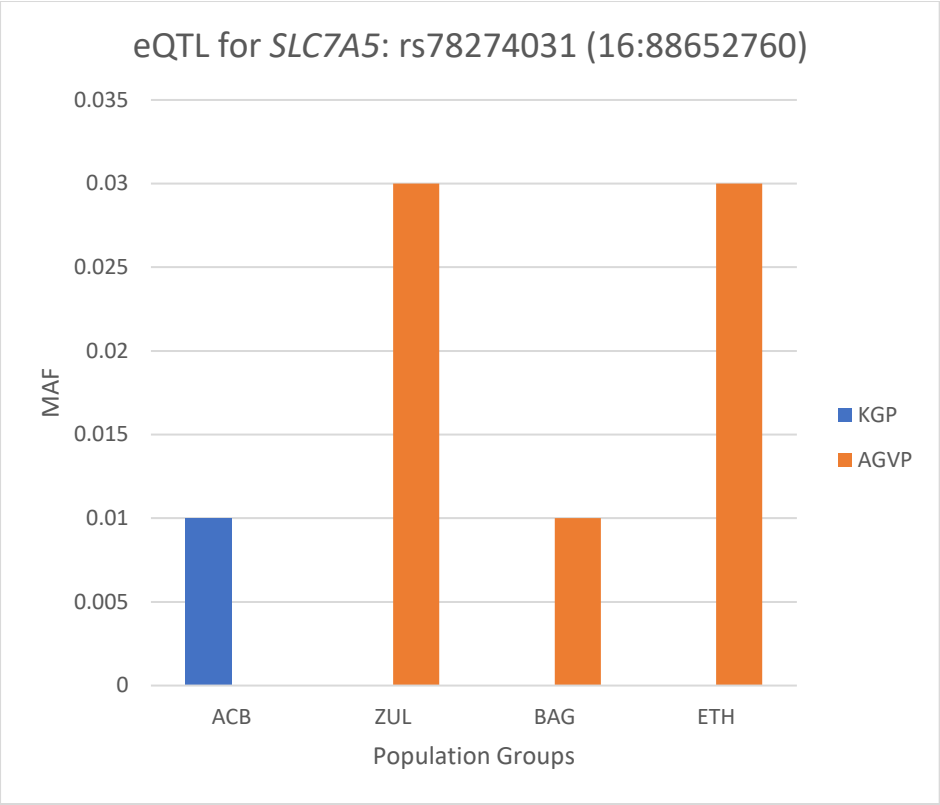


Figure 3.16: Allele frequencies of rs78274031 (16:88652760), eQTL for *SLC7A5* in African populations

Table 3.22: Allele frequencies of rs142290839 (15:57764596), eQTL for *ALDH1A2* in African populations

eQTL for *ALDH1A2*: rs142290839 (15:57764596)			
Dataset	Population	MAF	Sample size
AGVP	BAG	0.005	100
	ETH	0.013	120

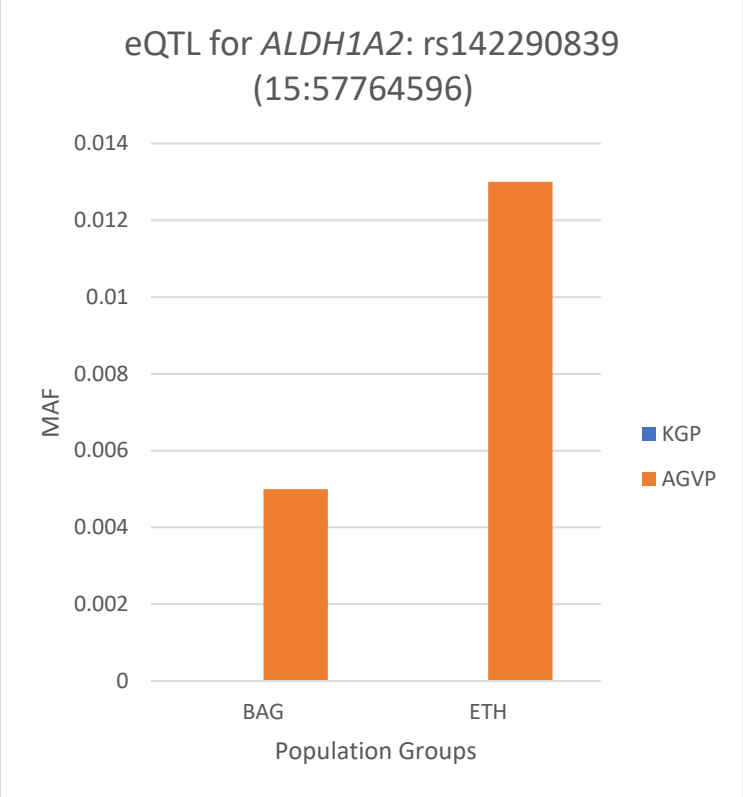


Figure 3.17: Allele frequencies of rs142290839 (15:57764596), eQTL for *ALDH1A2* in African populations

Table 3.23: Allele frequencies of rs117639661 (18:24994595), eQTL for *CHST9* in African populations

eQTL for *CHST9*: rs117639661 (18:24994595)			
Dataset	Population	MAF	Sample size
AGVP	ZUL	0.005	100

Table 3.24: Allele frequencies of rs151286679 (12:21013911), eQTL for *SLCO1A2* in African populations

eQTL for *SLCO1A2*: rs151286679 (12:21013911)			
Dataset	Population	MAF	Sample size
KGP	ASW	0.010	61
AGVP	ETH	0.004	120

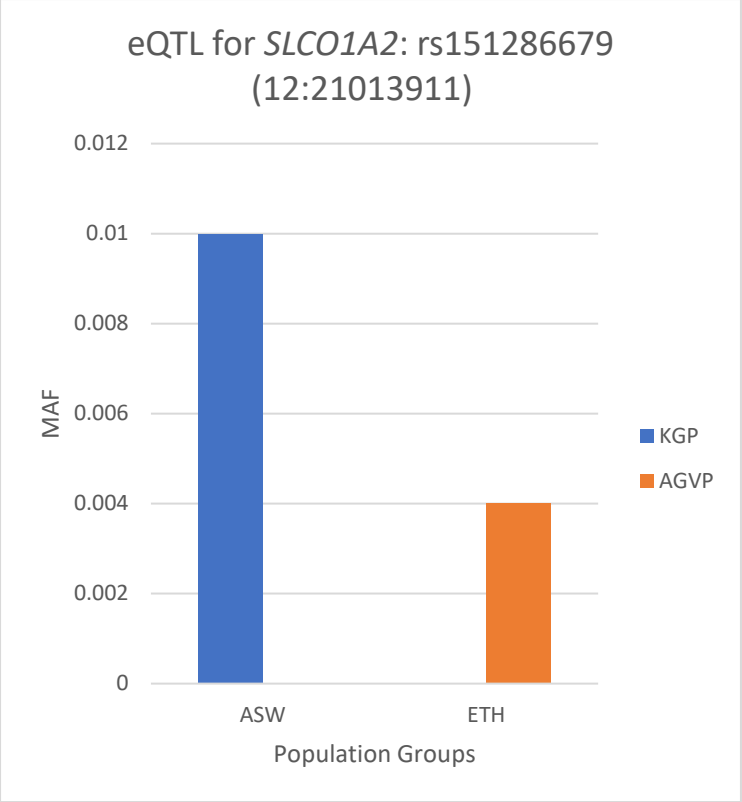


Figure 3.18: Allele frequencies of rs151286679 (12:21013911), eQTL for *SLCO1A2* in African populations

Table 3.25: Allele frequencies of rs74874632 (16:47510672), eQTL for *ABCC11* in African populations

eQTL for *ABCC11*: rs74874632 (16:47510672)			
Dataset	Population	MAF	Sample size
AGVP	ETH	0.004	120

Table 3.26: Allele frequencies of rs542683997 (7:150756557), eQTL for *ABCB8* in African populations

eQTL for *ABCB8*: rs542683997 (7:150756557)			
Dataset	Population	MAF	Sample size
KGP	ACB	0.005	96

Table 3.27: Allele frequencies of rs551125507 (21:36743550), eQTL for *CBR3* in African populations

eQTL for *CBR3*: rs551125507 (21:36743550)			
Dataset	Population	MAF	Sample size
KGP	ACB	0.005	96

The *cis*-eQTLs for *ALDH1A2, CHST9, SLCO1A2, ABCC11, ABCB8* and *CBR3* are very rare and only found in selected African populations. *Cis*-eQTLs for *ABCC3*, *SLC22A17*, *SLC7A5* and *ALDH1A2* are rare but found in more African populations than the *cis*-eQTLs for *ALDH1A2, CHST9, SLCO1A2, ABCC11, ABCB8* and *CBR3*. The *cis*-eQTLs of *SOD2*, *CYP26A1* and *SLC13A3* are found in all the African populations that we looked at. However, the *cis*-eQTLs for *SOD2* and *CYP26A1* are more common than that of SLC*13A3*.

Table 3.28 and Figure 3.19 shows a comparison of the allele frequencies of the variant allele of the *cis*-eQTLs in the five major population groups of the KGP, namely AFR (African), AMR (American), EAS (East Asian), EUR (European) and SAS (South Asian). The *cis*-eQTL for *UGT1A1*, *SOD2*, *CYP26A1* and *SLC13A3* are more frequently found in Africans than in Europeans.

Table 3.28: Allele frequencies of the *cis*-eQTLs across major populations of the KGP

***cis*-eQTL**	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**
rs28967009, eQTL for *UGT1A1*	0.05	0.07	0.03	0.01	0.04
rs114045355, eQTL for *DPYD*	0	0	NA	0.01	0.02
rs139101291, eQTL for *SOD2*	0.28	0.06	0.08	0.03	0.11
rs11817295, eQTL for *CYP26A1*	0.20	0.04	0.10	0.03	0.13
rs191131632, eQTL for *SLC13A3*	0.04	0	NA	NA	NA
rs146347791, eQTL for *ABCC3*	0.01	0	NA	0.01	0
rs140088214, eQTL for *SLC22A17*	0	0.17	NA	0	0
rs78274031, eQTL for *SLC7A5*	0	NA	0.11	0.01	0.03
rs142290839, eQTL for *ALDH1A2*	NA	0.01	0	0	0.02
rs117639661, eQTL for *CHST9*	NA	NA	NA	0.01	0.01
rs151286679, eQTL for *SLCO1A2*	0	0	NA	0.02	0
rs74874632, eQTL for *ABCC11*	NA	NA	0.04	0	NA
rs542683997, eQTL for *ABCB8*	0	0	NA	0	NA
rs551125507, eQTL for *CBR3*	0	0	NA	0	NA

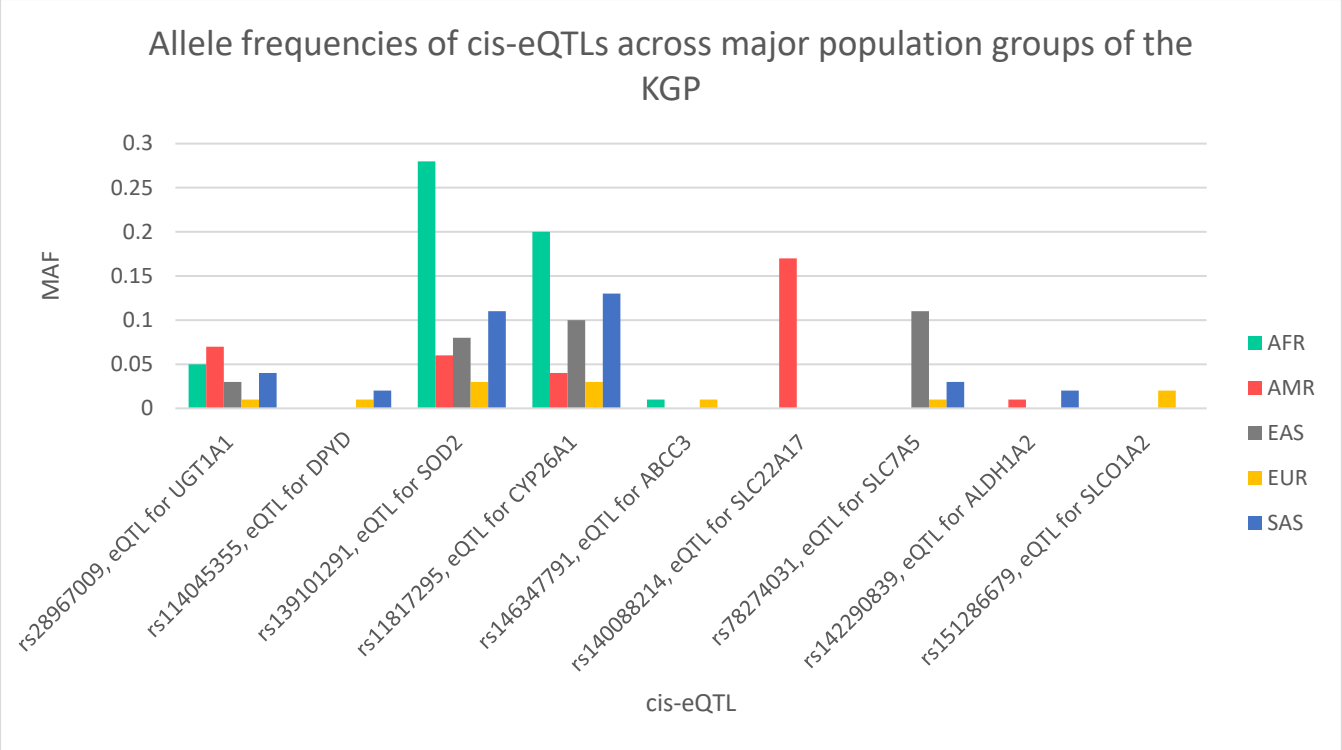


Figure 3.19: Allele frequencies of the *cis*-eQTLs across major populations of the KGP

3.10 EFFECT OF IMPUTATION AND SAMPLE SIZE ON NUMBER OF eQTLs IDENTIFIED

Dataset A with 145 samples generated 33 *cis*-eQTLs and 47922 *trans*-eQTLs (Table 3.3). Dataset B with 136 samples generated 36 *cis*-eQTLs and 46921 *trans*-eQTLs (Table 3.3). It was expected that a bigger dataset would generate more eQTLs compared to a smaller dataset. Therefore, it was quite surprising that dataset A with nine more samples than dataset B produced three fewer *cis*-eQTLs. Imputed dataset A with 145 samples generated more eQTLs (25962 *cis*-eQTLs and 38105386 *trans*-eQTLs) than imputed dataset B with 136 samples (25275 *cis*-eQTLs and 35837751 *trans*-eQTLs) as expected. Table 3.3 clearly shows that imputation drastically increases the number of *cis*- and *trans*-eQTLs generated.

3.11 eQTLs FOR CORE ADME GENES AND THEIR ASSOCIATION WITH DRUG RESPONSE-RELATED PHENOTYPES IN THE UK BIOBANK

The *cis*-eQTLs identified for the core ADME gene were searched for drug response related phenotypes in the UK Biobank using http://big.stats.ox.ac.uk/. Only eQTLs for *DPYD* and

UGT1A1, rs114045355 (Figure 3.20) and rs28967009 (Figure 3.21) respectively, were found to be associated with drug response related phenotypes.

The top hits for rs114045355 (*DPYD*) from Figure 3.20 are special screening examination for infectious and parasitic diseases, dorzolamide + timolol (used for treatment of high blood pressure inside of the eye), atrial fibrillation and flutter and securon 40 mg tablet (a calcium-channel blocker that is used for abnormal heart rhythms, high blood pressure and angina). The top hits for rs28967009 *(UGT1A1*) from Figure 3.20 are underlying primary cause of death for alcoholic liver disease, oramorph 10mg/5ml oral solution (liquid morphine used to treat moderate to severe pain), underlying primary cause of death for gastrointestinal haemorrhage and fractured wrist.

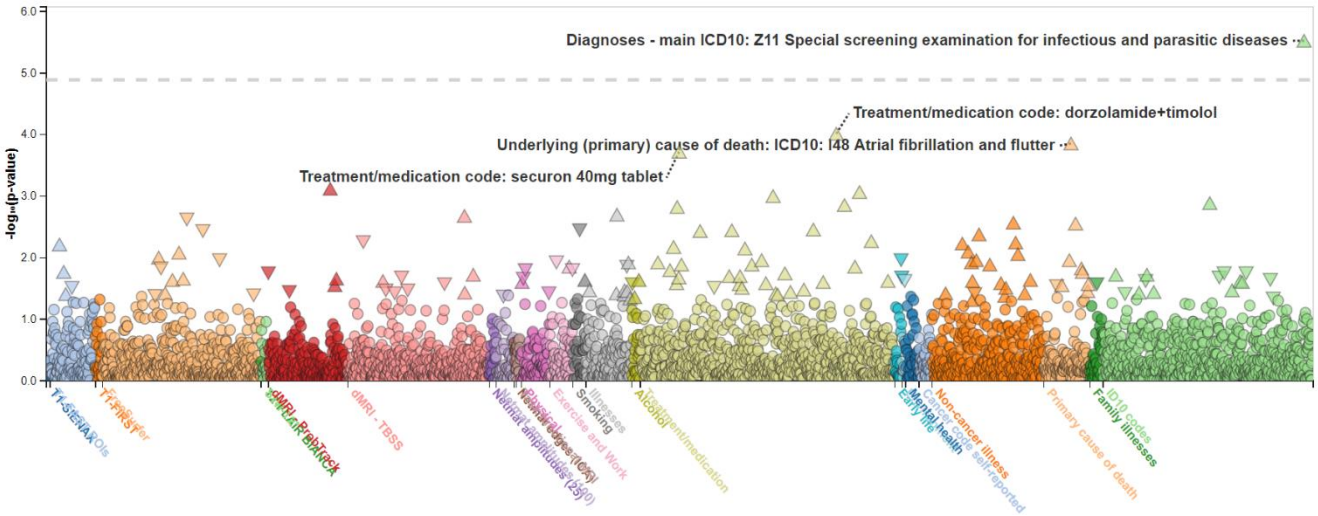


Figure 3.20: *Cis*-eQTL for *DPYD* (rs114045355) and its associated drug response related phenotype from UK Biobank

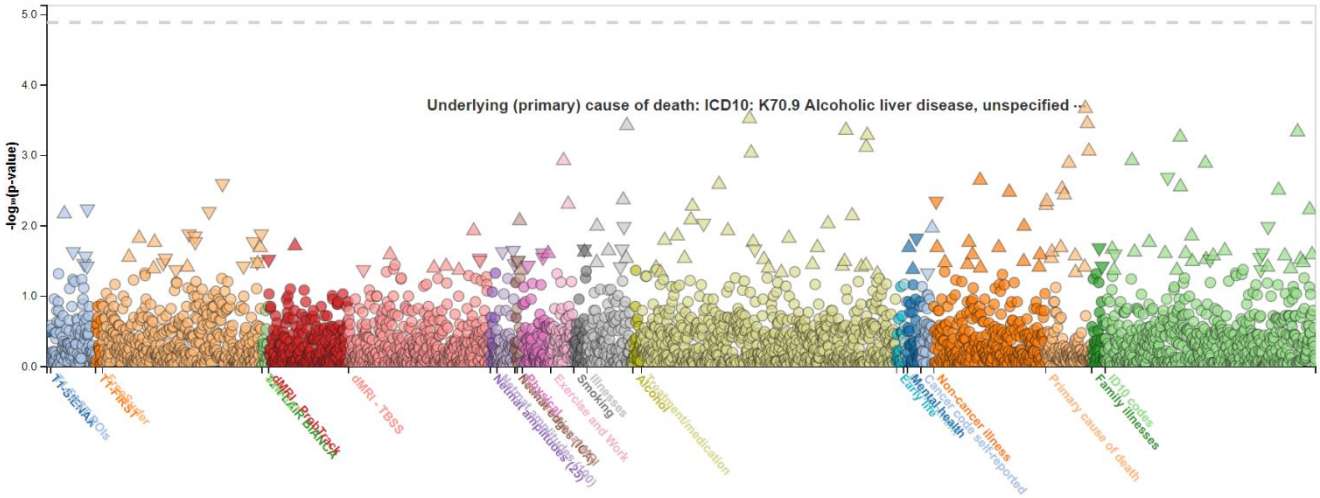


Figure 3.21: C*is*-eQTL for *UGT1A1* (rs28967009) and its associated drug response related phenotype from UK Biobank

3.12 ANNOTATION OF CORE ADME GENE eQTLs USING REGULOMEDB AND VEP

Cis-eQTLs identified for the core ADME genes were annotated using RegulomeDB and VEP. Table 3.29 shows the results obtained from RegulomeDB. *Cis*-eQTLs of *UGT1A1* and *CYP3A5* have a rank of 5 that corelates to 'Transcription factor (TF) binding or DNase peak'. *Cis*-eQTL for *CYP2C19* has a rank of 3a that corelates to 'TF binding + any motif + DNase peak'. *Cis*-eQTL for SULT1A1 has a rank of 6 that corelates to a 'Motif hit'. *Cis*-eQTLs of *DPYD* and *UGT2B17* have a rank of 7 that corelates to 'Other'.

Table 3.29: RegulomeDB annotation for *cis*-eQTLs of core ADME genes

***Cis*-eQTL**	**Chromatin state**	**Biosample**	**Rank**	**Score**
rs933765367, *cis*-eQTL for *CYP2C19*	Active TSS	HepG2	3a	0.56761
	Enhancers	liver		
rs143210517, *cis-eQTL* for *CYP3A5*	Enhancers	HepG2	5	0
	Quiescent/Low	liver		
rs114045355, *cis*-eQTL for *DPYD*	Quiescent/Low	HepG2	7	0.18412
	Quiescent/Low	liver		
rs28967009, *cis*-eQTL for UGT1A1	Strong transcription	liver	5	0.13454
	Quiescent/Low	HepG2		
rs917897719, *cis*-eQTL for UGT2B17	Repressed PolyComb	liver	7	0.18412
	Quiescent/Low	HepG2		
rs1000598999, *cis*-eQTL for SULT1A1	Repressed PolyComb	liver	6	0.77072
	Repressed PolyComb	HepG2		

Table 3.30 shows the results obtained from VEP. *Cis*-eQTLs for *CYP2C19*, *UGT2B17* and *SULT1A1* were found to be intron variants. *Cis*-eQTL for *CYP3A5* was found to be an upstream synonymous variant. *Cis*-eQTL for *DPYD* was found to be an intergenic variant and that of *UGT1A1* was found to be a downstream gene variant. *Cis*-eQTL for *CYP2C19* was identified to be in *SORBS1*. *Cis*-eQTL for *CYP3A5* was identified to be in *NYAP1*. *Cis*-eQTL for *UGT1A1* was identified to be in *UGT1A11P*. *Cis*-eQTL for *UGT2B17* was identified to be in *STAP1*. *Cis*-eQTL for *SULT1A1* was identified to be in *GSG1L*.

Table 3.30: VEP annotation for *cis*-eQTLs of core ADME genes

***Cis*-eQTL**	**Consequence**	**Gene**
rs933765367, *cis*-eQTL for *CYP2C19*	intron variant	*SORBS1*
rs143210517, *cis-eQTL* for *CYP3A5*	synonymous variant; upstream gene variant	*NYAP1*
rs114045355, *cis*-eQTL for *DPYD*	intergenic variant	
rs28967009, *cis*-eQTL for UGT1A1	downstream gene variant	*UGT1A11P*
rs917897719, *cis*-eQTL for UGT2B17	intron variant	*STAP1*
rs1000598999, *cis*-eQTL for SULT1A1	intron variant	*GSG1L*

3.13 ANNOTATION OF *cis*-eQTLs OF CORE ADME GENES

Drug label annotations and variant annotations were searched for each of the six core ADME genes, for which eQTLs were identified, in PharmGKB. No drug label annotations were found for *SULT1A1* and *UGT2B17*. Drug label annotations for the other four core ADME genes *CYP2C19, DPYD, CYP3A5* and *UGT1A1* are listed in Table 3.31 - Table 3.34. The following abbreviations have been used in tables numbered Table 3.31 through to Table 3.34: A- Actionable PGx; I- Informative PGx; Tr- Testing recommended; TR- Testing required. Variant annotations associated with the core ADME genes are not listed here as they are quite extensive. These can be accessed in PharmGKB.

Table 3.31: Drug label annotations for *CYP2C19*

Drug	Condition	Summary	US FDA	EMA	PMDA	HCSC	Swissmedic
Brivaracetam	Anticonvulsant	Brivaracetam is found in increased levels in CYP2C19 IMs or PMs. CYP2C19 (PMs) may require a dose reduction.	A	A		A	A
Carisoprodol	Skeletal muscle relaxant	Caution must be taken in administering Carisoprodol for patients with reduced CYP2C19 activity.	A				
Citalopram	Antidepressant	In CYP2C19 slow metabolisers, starting dose of 10 mg daily is recommended. Depending on patient response, the dose may be increased to a maximum of 20 mg per day.	A			A	A
Clobazam	Adjunctive treatment of seizures associated with Lennox-Gastaut syndrome in patients 2 years of age or older	An initial dose of 5 mg/day of clobazam must be prescribed in CYP2C19 PMs. The dose must be increased slowly based on weight.	A				
Clopidogrel	Platelet inhibitor	Clopidogrel has a reduced efficacy in CYP2C19 slow metabolisers. Slow metabolisers have reduced drug efficacy and are at higher risk for experiencing cardiovascular events. An alternative treatment must be considered in these patients or the use of higher doses of clopidogrel in slow metabolisers should be considered.	A	A	A	A	A

Dexlansoprazole	Reduces stomach acid/Protein pump inhibitor	Systemic exposure of the drug is higher in CYP2C19 IMs and PMs compared to EMs. CYP2C19 metaboliser phenotype can also affect the nature of some drug-drug interactions involving dexlansoprazole.	A			A	A
Diazepam	Breakthrough seizures	Diazepam is metabolised by CYP2C19 and CYP3A4. CYP2C19 and CYP3A4 genetic variants result in inter-individual variation in clearance of the drug.	A				
Doxepin	Antidepressant	CYP2D6 and CYP2C19 PMs have higher than expected plasma levels of the drug when given typical doses.	A				
Escitalopram	Treatment of acute and maintenance treatment of major depressive disorder in adults and adolescents aged 12-17 years and for acute treatment of generalized anxiety disorder	Dosing CYP2C19 slow metabolisers must be treated carefully. In CYP2C19 PMs, the maximum dose recommended is 10 mg due to an increased risk for developing QT prolongation and other side effects from the increased plasma concentrations.	A		A		A
Esomeprazole	Proton pump inhibitor	CYP2C19 PMs have higher plasma concentrations of the drug compared to EMs. Esomeprazole dose adjustment based on CYP2C19 genotype is not necessary.	A	A		A	

Flibanserin	Treatment of premenopausal women with acquired, generalized hypoactive sexual desire disorder (HSDD)	CYP2C19 PMs have increased exposure to the flibanserin as compared to EMs.	A				
Omeprazole	Reduces stomach acid/Protein pump inhibitor	CYP2C19 PMs have a greater systemic exposure to omeprazole, followed by IMs and EMs. CYP2C19 metaboliser status has no implication on omeprazole dosing.	A		A	A	A
Pantoprazole	Reduces stomach acid/Protein pump inhibitor	CYP2C19 PMs exhibited lower clearance of the drug as compared to EMs. Dose reduction must be considered in these patients. Dosage adjustment is not needed in adult CYP2C19 PMs.	A				A
Rabeprazole	Reduces stomach acid/Protein pump inhibitor	Gastric acid suppression was higher in CYP2C19 PMs compared to EMs, due to higher rabeprazole levels in PMs.	A		A	A	A
Voriconazole	Anti-fungal	Voriconazole levels are increased in patients who are CYP2C19 IMs or PMs. Due to interaction with other drugs that are substrates, activators or inhibitors of CYP3A4, CYP2C9 or CYP2C19 warning information regarding the co-administration of drugs is provided.	A	A	A	A	A

Atazanavir	Antiretroviral	Voriconazole and ritonavir must not be administered with atazanavir. In slow CYP2C19 metabolisers, close clinical monitoring for a loss of voriconazole and atazanavir efficacy is recommended. In slow metabolisers of CYP2C19 voriconazole exposure expected to be increased and atazanavir exposure is found to be decreased.		A	A	A	
Ticagrelor	Antiplatelet	Patients with slow CYP2C19 metabolism showed increased non-coronary artery by-pass grafting major bleeding when treated with ticagrelor compared to clopidogrel. Thrombotic cardiovascular events was not found to be dependent on CYP2C19 loss of function status.	A	A			
Lansoprazole	Reduces stomach acid/proton pump inhibitor		A				A
Moclobemide	Antidepressant	Moclobemide metabolism may be affected in genetically induced or drug-induced slow metabolisers. Caution is recommended when using moclobemide with drugs that are metabolised by CYP2C19 as moclobemide inhibits this enzyme.					A

Lacosamide	Partial-onset seizures	FDA label states that plasma concentrations of the lacosamide were reduced CYP2C19 PMs compared to EMs. There are no clinically relevant differences in lacosamide pharmacokinetics between PMs and EMs. EMA drug label says that no clinically relevant difference in lacosamide exposure was seen when comparing CYP2C19 extensive metabolisers and CYP2C19 poor metabolisers.	A	A			
Nelfinavir	Protease inhibitor used to treat human immunodeficiency virus (HIV)	Co-administration of nelfinavir and drugs that induce CYP3A or CYP2C19 may decrease nelfinavir plasma concentrations and reduce its therapeutic effect. Co-administration of nelfinavir and drugs that inhibit CYP3A or CYP2C19 may increase nelfinavir plasma concentrations.	A				
Prasugrel	Platelet inhibitor	Contains information regarding a lack of effect of polymorphisms in the *CYP3A5*, *CYP2B6*, *CYP2C9* and *CYP2C19* on prasugrel pharmacokinetics.	A	A			A

Summary obtained from https://www.pharmgkb.org/gene/PA124/labelAnnotation

Table 3.32: Drug label annotations for *DPYD*

Drug	Condition	Summary	US FDA	EMA	PMDA	HCSC	Swissmedic
Capecitabine	Anti-cancer	Dihydropyrimidine dehydrogenase (DPD) deficiency increases the risk of fatal ADRs. It is contraindicated in patients with DPD deficiency. No drug dose is safe in patients with no DPD activity. Patients with partial DPD activity may be given the drug at a reduced dose if the benefit of the drug outweighs the risks. The EMA recommends genotyping of DPYD identify patients at increased risk for severe toxicity.	A	Tr	A	A	A
Fluorouracil	Anti-cancer	DPD deficiency puts patients at increased risk for severe toxicity. Swissmedic requires DPYD testing for fluorouracil in patients who have also been treated with brivudine or other nucleoside analogues. Swissmedic recommends genotyping DPYD prior to treatment.	A		A	A	TR
Flucytosine	Anti-fungal	Patients with DPD deficiency are at hight risk for drug toxicity if nucleoside analogues have been given. DPYD testing is recommended.					Tr

Summary obtained from https://www.pharmgkb.org/gene/PA145/labelAnnotation

Table 3.33: Drug label annotations for *CYP3A5*

Drug	Condition	Summary	US FDA	EMA	PMDA	HCSC	Swissmedic
Maraviroc	Antiretroviral	CYP3A5 genotypes may affect the exposure to maraviroc.					A
Prasugrel	Platelet inhibitor	Contains information regarding a lack of effect of genetic variants in the CYP3A5, CYP2B6, CYP2C9 and CYP2C19 genes on prasugrel pharmacokinetics.	I	I			I

Summary obtained from https://www.pharmgkb.org/gene/PA131/labelAnnotation

Table 3.34: Drug label annotations for *UGT1A1*

Drug	Condition	Summary	US FDA	EMA	PMDA	HCSC	Swissmedic
Belinostat	Anti-cancer	Patients homozygous for UGT1A1*28 allele must minimize dose to limit toxicities. Starting dose must be lowered to 750 mg/m2 in such patients.	A				
Dolutegravir	Anti-retroviral	UGT1A1 PMs have decreased clearance and increased AUC as compared to normal UGT1A1 metaboliser. The label does not mention specific UGT1A1 variants or genetic testing. Swissmedic states that there is no evidence that gene variants of UGT1A1, CYP3A4, CYP3A5 and NR1I2 have clinically relevant effects.	A	A			I

Irinotecan	Anti-cancer	Starting dose of irinotecan must be reduced for patients homozygous for *UGT1A1*28* due to risk for hematologic toxicity. PMDA recommends that patients should be selected for treatment based on the stage of their cancer, general condition and UGT1A1 polymorphism.	A	A	Tr	A	A
Nilotinib	Anti-cancer	Individuals with the UGT1A1*28 genotype (TA)7/(TA)7 (rs8175347) are at an increased risk of hyperbilirubinemia when taking nilotinib.	A			A	
Erlotinib	Anti-cancer	Patients with low expression levels of UGT1A1 or genetic glucuronidation disorders such as Gilbert's disease may have increased concentrations of bilirubin and should be treated with caution.		A			
Indacaterol	Bronchodilator	UGT1A1 (TA)7(TA)7 genotype is associated with increased exposure to indacaterol.	I		A		I

Summary obtained from https://www.pharmgkb.org/gene/PA420/labelAnnotation

3.14 ANNOTATION OF *cis*-eQTLs OF EXTENDED ADME GENES

Variant annotations found for extended ADME genes involved in the metabolism of drugs commonly prescribed for HIV and AIDS, tuberculosis, diabetes, hypertension and cardiovascular diseases, and over-the-counter medications and prescription medications for pain are listed below. The above-mentioned conditions are listed in the top ten causes of mortality in South Africa published by the Department of Statistics, South Africa in 2017.

Table 3.35: Variant annotations for *UGT1A6*

UGT1A6	
Variants	**Effect of the variant**
*UGT1A6*2a*	*UGT1A6 *2a* is associated with increased activity of UGT1A6 when assayed with acetaminophen in HEK cells as compared to *UGT1A6 *1a*.
*UGT1A6*2a*	*UGT1A6 *2a* is associated with increased metabolism of acetaminophen in healthy individuals as compared to *UGT1A6 *1a*.
*UGT1A6*2a*	*UGT1A6 *2a* is associated with increased clearance of acetaminophen in people with beta-Thalassemia as compared to *UGT1A6 *1a*.

Summary obtained from https://www.pharmgkb.org/gene/PA37181/variantAnnotation

Table 3.36: Variant annotations for *ABCC3*

ABCC3	
Variants	**Effect of the variant**
rs1051640	Allele G is associated with increased risk of Ototoxicity when treated with cisplatin in children with Neoplasms as compared to allele A. Other studies have contradicting results
rs4148416	Allele T is associated with increased risk of death when treated with cisplatin, cyclophosphamide, doxorubicin, methotrexate and vincristine in people with Osteosarcoma as compared to allele C.

Summary obtained from https://www.pharmgkb.org/gene/PA376/variantAnnotation

Table 3.37: Variant annotations for *ABCC4*

ABCC4	
Variants	**Effect of the variant**
rs1751034	Genotype TT is associated with increased tenofovir renal clearance, and lower AUC when treated with tenofovir as compared to genotypes CC + CT.
rs1059751	Allele G is associated with increased likelihood of Kidney Diseases when treated with tenofovir as compared to allele A.
rs3742106	Allele C is associated with increased concentrations of tenofovir as compared to allele A.
rs9561778	Allele T is associated with increased risk of ADR when treated with cyclophosphamide, doxorubicin and fluorouracil in people with Breast Neoplasms as compared to allele G.
rs9561778	Allele T is associated with increased risk of gastrointestinal toxicity when treated with cyclophosphamide in people with Breast Neoplasms as compared to allele G.
rs9561778	Allele T is associated with increased severity of Neutropenia when treated with cyclophosphamide in people with Breast Neoplasms as compared to allele G.
rs16950650	Allele T is associated with decreased overall survival when treated with cisplatin and irinotecan in people with Small Cell Carcinoma as compared to allele C.
rs1059751	Allele G is associated with increased likelihood of kidney diseases when treated with tenofovir in people with HIV Infections as compared to allele A.

Summary obtained from https://www.pharmgkb.org/gene/PA397/variantAnnotation

Table 3.38: Variant annotations for *XDH*

XDH	
Variants	**Effect of the variant**
rs1429376	Genotype AA is associated with increased risk of developing noncirrhotic portal hypertension when treated with didanosine in people with HIV Infections as compared to genotypes AC + CC.
rs1594160	Genotype AA is associated with increased risk of developing noncirrhotic portal hypertension when treated with didanosine in people with HIV Infections as compared to genotypes AC + CC.

Summary obtained from https://www.pharmgkb.org/gene/PA37404/variantAnnotation

Table 3.39: Variant annotations for *CES1*

CES1	
Variants	**Effect of the variant**
rs2244613	Allele G is associated with increased risk of Drug Toxicity when treated with capecitabine in people with Neoplasms as compared to allele T.
rs2244614	Allele G is associated with increased risk of Drug Toxicity when treated with capecitabine in people with Neoplasms as compared to allele A.
rs3217164	Allele G is associated with increased risk of Drug Toxicity when treated with capecitabine in people with Neoplasms as compared to allele del

Summary obtained from https://www.pharmgkb.org/gene/PA107/variantAnnotation

CHAPTER 4 DISCUSSION

The stringent quality control process led to a significant attrition of usable data. The genotyped variants were reduced to 297,585 following QC of the initial 318,237 SNPs. Four samples were removed because of high missingness, and seven samples were identified as population outliers from the genotype dataset. Six samples were identified as outliers from the gene expression dataset. After mapping the samples that failed QC from both genotype and gene expression datasets, a total of thirteen samples had to be excluded from further analysis. The final dataset included 136 samples.

Local *cis*-eQTLs were identified that were located within 1,000,000 base pairs (1 Mb pair) upstream or downstream of the TSS of a gene. A linear model was used for the eQTL analysis. A variant was regarded as a significant eQTL for a given gene only if the p-value and FDR were less than 0.05. Age and sex were covariates in the eQTL analysis.

Cis-eQTLs were identified for 6 core ADME genes, namely *CYP2C19*, *CYP3A5*, *DPYD*, *SULT1A1*, *UGT1A1* and *UGT2B17*. These genes were searched in PharmGKB to identify a list of drugs associated with them.

4.1 CORE ADME GENES AND THEIR ASSOCIATED eQTLs

The effect the identified *cis*-eQTL has on the drugs metabolised by the core ADME genes is discussed below. For the purpose of this study, I only focussed on drugs listed in the drug label annotation section associated with each of the six core ADME genes in PharmGKB and on their variant annotations. The variant annotations of the core ADME genes can be found on PharmGKB.

4.1.1 CYP2C19

CYP2C19 encodes for a phase I drug metabolising enzyme, cytochrome P450 2C19. The chromosomal location of *CYP2C19* is 10q23.33. *CYP2C19* is highly polymorphic.

Polymorphisms in *CYP2C19* are associated with isozymes with altered enzyme activity (Cleveland Clinic Laboratories, 2013). The *CYP2C19*1* allele corresponds to fully functional metabolism and is referred to as the wild type allele (Mirabbasi *et al.*, 2017). *CYP2C19 *2*, **3*, **4*, **5*, **6*, **7*, and **8* alleles are associated with non-functional CYP2C19 activity and *CYP2C19 *9* and **10* are associated with reduced functional CYP2C19 activity (Mirabbasi *et al.*, 2017). Increased enzyme activity is observed in *CYP2C19*17* allele carriers because of increased gene transcription (Mirabbasi *et al.*, 2017). Based on which CYP2C19 alleles an individual has, they can be classified as UM, NM, IM, and PM (Cleveland Clinic Laboratories, 2013).

This study identified rs933765367 (10:97236373) as an eQTL for *CYP2C19*. The p-value (5.64E-08) and FDR (0.0009) for this association are significant. The beta value for this eQTL-gene association is 3238. This means that the variant allele C of the eQTL/eSNP is associated with significant increase (3238 times on average) in the expression of *CYP2C19.* The effect of the *cis*-eQTL on *CYP2C19* expression is illustrated in Figure 4.1.

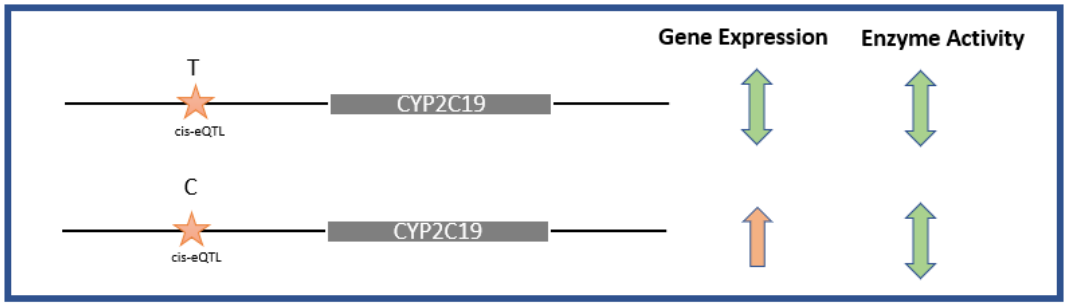


Figure 4.1: The effect of the variant allele C of the *cis*-eQTL, rs933765367, on CYP2C19 enzyme activity. The reference allele, T, is associated with normal (usual) *CYP2C19* expression whereas the variant allele, C, is associated with increased expression of *CYP2C19*

The effect of the *cis*-eQTL, rs933765367, on the metabolism of the following drugs are discussed in detail in the sections below:

1. Clopidogrel, an antiplatelet medication;
2. Omeprazole, esomeprazole, dexlansoprazole and rabeprazole, proton pump inhibitors; and

3. Citalopram and escitalopram, anti-depressants.

4.1.1.1 Clopidogrel and CYP2C19

Clopidogrel is an antiplatelet medication that is prescribed to prevent heart attack and stroke (Sangkuhl *et al.*, 2010; Mirabbasi *et al.*, 2017). It is prescribed for people who are at high risk of heart disease and stroke and who have recently had a heart attack or stroke. It is also prescribed for poor blood circulation in the legs, a condition called peripheral arterial disease. The main function of clopidogrel is to prevent the formation of dangerous clots. Sometimes, it is also prescribed as combination therapy with aspirin for people who have had heart attacks and following coronary artery stent placement.

Clopidogrel is a pro-drug that is converted to its active metabolite by two sequential oxidative steps (Sangkuhl *et al.*, 2010; Hulot *et al.*, 2012). In the first step, clopidogrel is converted to 2-oxo-clopidogrel by CYP3A4, CYP1A2, CYP2B6, and CYP2C19. CYP3A4, CYP2C9, CYP2C19 and CYP2B6 then convert 2-oxo-clopidogrel to the active metabolite. *CYP2C19*, thus plays a significant role in both the oxidative steps.

Depending on *CYP2C19* genotype, the pharmacokinetics of the active and intermediate metabolite and their antiplatelet effect will vary (Hulot *et al.*, 2012; Mirabbasi *et al.*, 2017). Doctors therefore recommend testing the genotype of *CYP2C19* before prescribing clopidogrel. The US FDA, EMA, PMDA, HCSC and Swissmedic indicate in their drug label that CYP2C19 slow metabolisers have reduced drug efficacy when compared to patients with normal CYP2C19 function and are at higher risk for experiencing cardiovascular events. In poor metabolisers, at the recommended dose, clopidogrel forms 70% less active metabolite and has a much smaller effect on inhibition of platelet aggregation. CYP2C19 poor metabolisers with acute coronary syndrome or undergoing percutaneous coronary intervention may be at an increased risk for experiencing cardiovascular related adverse drug events like myocardial infarction and stent thrombosis when clopidogrel is taken at the standard dose recommended as compared to CYP2C19 extensive metabolisers. The FDA,

hence, recommends an alternate platelet inhibitor in CYP2C19 poor metabolisers. It has been found in CYP2C19 poor metabolisers that a higher dosage of the drug increases the antiplatelet effect. Thus, the use of clopidogrel in higher doses must be considered in poor metabolisers.

It is predicted that the variant allele C of the *cis*-eQTL rs933765367 may significantly increase the expression of *CYP2C1* implying that, if an individual carries the fully functional CYP2C19 genotype (CYP2C19 EMs and UMs) in addition to rs933765367C, clopidogrel is likely to be converted to its active intermediate much faster because of increased expression levels of CYP2C19. This might result in increased levels of the active metabolite in circulation which can cause adverse drug reactions. Therefore, it is postulated that the drug dosage of clopidogrel should be lowered in individuals with the fully functional *CY2C19* genotype in the presence of the C allele of the eQTL or another drug must be recommended. If rs933765367C is found in CYP2C19 slow metabolisers (PMs and IMs), though the quantity of the enzyme with reduced activity increases because of increased *CYP2C19* expression, in order to reach a therapeutic concentration of the drug in circulation, it is suggested that clopidogrel should be prescribed in higher doses. Despite increasing the drug dose, if adequate drug efficacy is not achieved, an alternate drug must be prescribed. It is however premature to suggest changes in dosing recommendations based on the predicted effects of the novel eQTL until there is further evidence from clinical studies relating to its drug outcome.

4.1.1.2 Omeprazole, esomeprazole, dexlansoprazole, rabeprazole and CYP2C19

Omeprazole, esomeprazole, dexlansoprazole and rabeprazole are proton pump inhibitors (PPIs) that are used to treat gastrointestinal conditions like gastro-oesophageal reflux disease (GERD), and other acid-related disorders that are caused by high levels of acid in the stomach (El Rouby *et al.*, 2018). They work by reducing the amount of acid produced in the stomach. They are prescribed in combination with antibiotics to treat *Helicobacter pylori* (*H. pylori*) bacterial infections in the stomach. They are widely prescribed globally (Skrzydło-Radomańska *et al.*, 2015). They work by inhibiting the activity of the hydrogen/potassium

adenosine triphosphatase enzyme (H/K ATPase), also called the proton pump. Proton pumps are found in parietal cells of the gastric mucosa. They are involved in the secretion of hydrochloric acid (HCl). In the process of secreting HCL, ATP hydrolyses and H^+ ions are exchanged from the cytoplasm for K^+ ions. PPIs are pro-drugs that are activated in acidic environments (Hagymási *et al.*, 2011). After reaching the gastric parietal cells in the stomach that are activated by a meal, PPIs undergo protonation in the acidic pH environment, followed by conversion to sulphonamide. Sulphonamides are the active form of the drug that inhibit the activity of the proton pump. They are metabolised/inactivated in the liver by cytochrome P450 3A (CYP3A) and cytochrome P450 2C19 (CYP2C19) (Shin *et al.*, 2013; El Rouby *et al.*, 2018). The *CYP2C19*17* allele (increased CYP2C19 enzyme activity) enhances PPI clearance and results in poor PPI efficacy. The *CYP2C19*2* loss-of-function allele (no CYP2C19 enzyme activity) decreases the clearance of PPIs, resulting in more active PPI being available. This can either result in enhanced treatment or ADRs as the PPIs will remain in circulation for extended duration. There is a correlation in patients with low or absent CYP2C19 activity with increased drug efficacy, resulting in better overall outcomes, when treated with omeprazole. This is due to higher drug levels in patients. (Robinson *et al.*, 2003; Blume *et al.*, 2006; Tang *et al.*, 2013).

Dexlansoprazole, omeprazole (composed of equal amounts of the R and S enantiomers) and rabeprazole are delayed-release oral medications (Shin *et al.*, 2013). Delayed release medication is slowly released into the body over time. Esomeprazole, a pure S-isomer of omeprazole, can either be administered as a delayed release oral medication or as an intravenous (IV) medication that can only be administered by a health-care provider. Systemic exposure to PPIs varies with CYP2C19 metabolism status. It is highest in poor metabolisers followed by intermediate metabolisers, extensive metabolisers and then ultra-rapid metabolisers (Furuta *et al.*, 2005; Kuo *et al.*, 2014).

The US FDA and Swissmedic indicate in the drug label annotation for omeprazole that the risk of drug-drug interaction is significantly high in CYP2C19 slow metabolisers as *CYP2C19* is competitively inhibited by omeprazole (Blume *et al.*, 2006). The US FDA indicates in the drug

label annotation for esomeprazole in PharmGKB that CYP2C19 PMs have more drug in systemic circulation compared to NMs. The US FDA drug label for dexlansoprazole states that the systemic exposure of the drug is higher in CYP2C19 intermediate and poor metabolisers compared to extensive metabolisers. Dexlansoprazole is a substrate for CYP2C19. Therefore, co-administration of dexlansoprazole with other drugs that are also substrates for CYP2C19 can lead to toxic levels of those drugs, if their therapeutic index is narrow, as they will be inhibited by dexlansoprazole (Venkataramanan *et al.*, 1995; Imamura *et al.*, 2016). The US FDA, HCSC and Swissmedic makes note of this in their drug label annotations and PharmGKB has it labelled as an 'Actionable PGx'. The US FDA, PMDA, EMA and Swissmedic indicate in the drug label annotation of rabeprazole that the drug metabolism is slow in CYP2C19 PMs. Slow rabeprazole metabolism results in higher gastric acid suppression in CYP2C19 PMs than in CYP2C19 NMs because of higher rabeprazole plasma levels in PMs compared to NMs. Therefore, before prescribing the drug in poor metabolisers, doctors must adjust drug dosage levels to prevent complete suppression of stomach acid formation. CYP2C19 poor metabolisers will thus require a lower drug dose than the recommended rabeprazole dosage level.

Although it is premature to suggest changes in dosing recommendations based solely on the predicted effects of the novel eQTL until it is clinically and functionally validated, it is predicted that since the *cis*-eQTL rs933765367C, is associated with a significant increase in *CYP2C19* expression, the metabolism of the PPIs omeprazole, esomeprazole, dexlansoprazole and rabeprazole will occur faster because of increased enzyme levels due to increased gene expression. This implies that the dosage levels of the PPIs might have to be increased in CYP2C19 NMs and UMs with rs933765367C to achieve ideal stomach acid levels. However, in CYP2C19 PMs and IMs as the *cis*-eQTL rs933765367C increases the expression of the enzyme with reduced activity, the drug doses of the PPIs may need to be lowered to reduce the intensity and occurrence of ADRs. Ideally an alternate treatment regimen must be recommended to avoid ADRs associated with having the PPIs in circulation for extended periods of time in CYP2C19 PMs with rs933765367C.

4.1.1.3 Citalopram, escitalopram and CYP2C19

Citalopram is a chiral drug that is composed of an S-enantiomer and an R-enantiomer. The S-enantiomer, escitalopram, is therapeutically active and the R-enantiomer is therapeutically inactive. Citalopram and escitalopram are selective serotonin reuptake inhibitors (SSRI) prescribed for the treatment of acute major depressive disorder (MDD) in adults and adolescents aged 12-17 years. They are also used for the treatment of generalized anxiety disorder (GAD), obsessive compulsive disorder (OCD) and panic attacks. They work by helping to restore the balance of serotonin in the brain. This helps improve energy levels and feelings of well-being, decreases nervousness, and hence helps with anxiety. Escitalopram and citalopram are among the most prescribed antidepressants in paediatric patients (Czaja *et al.*, 2013). However, about 50% of these patients fail to respond effectively to the drug because they develop ADRs and have to discontinue treatment (Emslie *et al.*, 2009). Citalopram and escitalopram have fewer side-effects than most anti-depressants and hence are prescribed quite extensively. They are both oral medications than can be taken either in the form of a tablet or a liquid. They are primarily metabolised/cleared from the body by the cytochrome P450 2C19 enzyme and cytochrome P450 3A4 enzyme with minor involvement of cytochrome P450 2D6 enzyme (Sangkuhl *et al.*, 2011). Therefore, genetic polymorphisms of *CYP2C19* affect plasma levels of the drugs. The *CYP2C19*17* homozygous genotype associated with the ultra-rapid metaboliser phenotype has been found to be associated with increased escitalopram metabolism (Ohlsson Rosenborg *et al.*, 2008; Rudberg *et al.*, 2008). At equivalent drug doses, CYP2C19 ultra rapid metabolisers have lower drug serum concentrations compared to extensive CYP2C19 metabolisers. CYP2C19 PMs have higher serum concentrations of the drugs as they are metabolised very slowly by CYP2C19 (Chang *et al.*, 2014). This influence of CYP2C19 metabolic activity on plasma concentration of the drug is similar for both escitalopram and citalopram (Chang *et al.*, 2014). Thus, CYP2C19 rapid and extensive metabolisers are at a higher risk for treatment failure as they are cleared from the system faster, and CYP2C19 poor metabolisers are at increased risk of experiencing ADRs.

It is suggested that CYP2C19 poor metabolisers only use 61% of the standard recommended dose of citalopram (Mrazek *et al.*, 2011; Aldrich *et al.*, 2019). Therefore, when prescribing escitalopram in CYP2C19 poor metabolisers, dosage levels must be lowered accordingly. For

citalopram, in CYP2C19 PMs, for the first two weeks of treatment an initial daily dose of 10 mg and a maximum daily dose of 20 mg is recommended due to the risk for QT prolongation. For escitalopram, an initial dose of 5 mg per day for the first 2 weeks is recommended. Due to increased risk of developing QT prolongation and various other side effects from increased plasma concentrations of escitalopram in CYP2C19 poor metabolisers, a maximum dose of 10 mg is recommended. US FDA, PMDA and Swissmedic have drug labels for escitalopram where they emphasize the need for lowering drug dosage levels in CYP2C19 poor metabolisers.

In CYP2C19 extensive and ultra-rapid metabolisers with the *cis*-eQTL rs933765367C, it is postulated that citalopram and escitalopram will have to be prescribed in much higher doses to attain optimal drug effect as drug metabolism might occur faster due to increased expression of *CYP2C19*. However, in CYP2C19 intermediate and poor metabolisers with rs933765367C, both drugs may need to be administered in lower doses to lower the risk of ADRs. In CYP2C19 poor metabolisers with rs933765367C, it is suggested that either the drugs may need to be administered in much lower doses or an alternate treatment regime may be recommended. These suggestions on drug doses are made purely based on bioinformatic predictions and must be clinically and functionally validated.

4.1.1.4 Variant annotations for CYP2C19 and the effect of the eQTL

CYP2C19 metabolises several other drugs. Variant annotations of *CYP2C19* can be accessed at https://www.pharmgkb.org/gene/PA124/variantAnnotation. In breast cancer patients treated with cyclophosphamide, rs4244285A and rs12248560C are linked with increased risk of neutropenia and leukopenia respectively compared to alleles G and T respectively (Tecza *et al.*, 2018). If these patients also carry the *cis*-eQTL rs933765367C, it is hypothesised that the risk of developing neutropenia and leukopenia will increase considerably because of increased enzyme production as a result of increased *CYP2C19* expression. In such patients, cyclophosphamide may need to be administered under supervision to avoid extreme cases of ADRs. If the patients cannot tolerate the ADRs, treatment might need to be stopped immediately, and an alternate treatment regime might be suggested. Non-Hodgkin's lymphoma patients with *CYP2C19 *2* allele (non-functional *CYP2C19* enzyme activity) are

associated with decreased metabolism and decreased response to cyclophosphamide (Shu *et al.*, 2017). If they also carry the *cis*-eQTL rs933765367C, it is assumed that the expression of the *CYP2C19 *2* allele will increase significantly. Therefore, in patients who are poor metabolisers of CYP2C19, an alternate treatment regime might need to be recommended. Since allele A of rs4986893 is associated with a significant decrease in risk of maculopapular exanthema/skin rash when treated with ethambutol, isoniazid, pyrazinamide or rifampin in people with Tuberculosis compared to allele G (Richardson *et al.*, 2018), this combinatorial treatment will be ideal to treat tuberculosis in people with the *cis*-eQTL rs933765367C and rs4986893A. All the recommendations made with respect to changes in drug dose must be clinically and functionally validated since they are made based on bioinformatic predictions.

4.1.2 DPYD

The *DPYD* gene spans about 950kb on chromosome 1p22 and codes for the enzyme, dihydropyrimidine dehydrogenase (DPD). DPD is involved in breaking down fluoropyrimidine to its inactive metabolites. The CPIC lists 82 *DPYD* variants (Amstutz *et al.*, 2018). Polymorphisms in the *DPYD* gene cause DPD deficiency. Depending on the genotype of *DPYD,* individuals can be categorised as normal, intermediate and poor metabolisers (Henricks *et al.*, 2015).

Our analysis shows that rs114045355 (1:97464825) is a *cis*-eQTL for *DPYD* with a p-value of 8.27E-07, FDR of 0.008 and a beta-value of 325. This means that the variant allele T of the *cis*-eQTL rs114045355 increases the expression of *DPYD* significantly (325 times on average). The effect of this *cis*-eQTL on *DPYD* expression is illustrated in Figure 4.2. The effect of the *cis*-eQTL on the metabolism of the chemotherapeutics, capecitabine and fluorouracil is discussed in the section below.

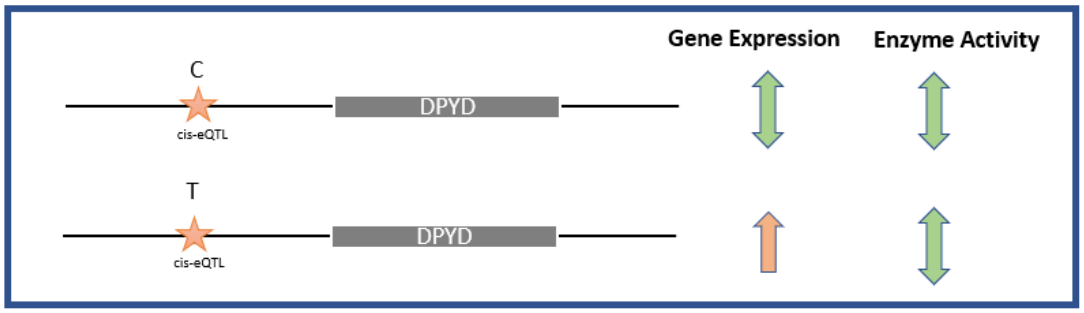


Figure 4.2: The effect of the variant allele T of the *cis*-eQTL, rs933765367, on DPYD enzyme activity. The reference allele, C, is associated with normal (usual) *DPYD* expression whereas the variant allele, T, is associated with increased expression of *DPYD*.

4.1.2.1 Capecitabine, fluorouracil and DPYD

Capecitabine is an oral anti-cancer/ anti-neoplastic/ cytotoxic chemotherapy drug that is used to treat several types of neoplasms including colorectal and breast neoplasms. It is a type of chemotherapy drug that is classified as an anti-metabolite. When cells incorporate anti-metabolites into their cellular metabolism, cell division is inhibited. They attack cells at specific phases in the cell-cycle and are therefore cell-cycle specific. Thus, its effect is exerted by preventing cancerous cells from dividing and repairing damaged DNA.

Capecitabine is a prodrug which is metabolised into a more common active form of the drug called fluorouracil (Loganayagam *et al.*, 2013; Jacobs *et al.*, 2019). It is absorbed through the intestine into the liver where it is first converted to 5'-deoxy-S-fluorocytidine (5'-DFCR) by carboxylesterase and then to 5'-deoxy-S-fluorouridine (5'-DFUR) by cytidine deaminase (Cyt D). The enzyme thymidine phosphorylase (TP) then converts 5'-DFUR to the active form of the drug, fluorouracil (5-FU) in both normal and tumour tissues. This mode of action increases tumour-targeting specificity thereby decreasing systemic drug exposure. As a drug, 5-FU is available both as an injection and a topical cream. However, these two formulations have completely different indications. Fluorouracil injection can either be used as a monotherapy or in combination with other cancer drugs for the treatment of various types of cancer like colorectal, breast, head and neck, anal and stomach cancer. Fluorouracil cream is used for the treatment of multiple actinic or solar keratoses of the face and anterior scalp.

The enzyme dihydropyrimidine dehydrogenase (DPD) is the primary enzyme involved in the degradation of capecitabine and 5-FU. Certain polymorphisms like *DPYD*2A*, c.1679T>G (*DPYD*13*), c.2846A>T, c.1236G>A in the *DPYD* locus can cause complete to near complete loss of DPD enzymatic activity (Jennings *et al.*, 2013). People with these mutations are at an increased risk for life-threatening/fatal toxicity because the drug remains in circulation for extended periods of time in the case of someone with reduced DPD activity and it cannot be removed from circulation in people with complete loss of DPD activity. For patients that are prescribed Capecitabine/5-FU treatment who show signs of acute early-onset or severe drug toxicity, the treatment must be immediately withheld or discontinued. Drug toxicity is an indication of near outright or outright absence of DPD activity. No drug dose of capecitabine/5-FU has been proven safe in patients with complete absence of DPD activity and therefore capecitabine/5-FU is contraindicated in such patients. The EMA highly recommends genotyping for certain alleles to identify patients at increased risk for capecitabine related severe toxicity. The Swiss drug label for fluorouracil requires *DPYD* testing in patients prior to 5-FU treatment.

Patients with certain heterozygous *DPYD* variants (including *DPYD*2A*, *DPYD*13*, c.2846A>T and c.1236G>A/HapB3 variants) have been shown to have increased risk of severe toxicity when treated with capecitabine/5-FU (Henricks *et al.*, 2017). In order to identify people at high risk for severe capecitabine/5-FU related drug toxicity, genotyping these alleles is highly recommended.

In patients with partial DPD deficiency, where benefits of capecitabine/5-FU treatment are thought to outweigh potential risks, capecitabine/5-FU treatment must be administered with extreme caution. The dose of the drug must be frequently monitored and adjusted depending on the extent of toxicity. In such patients, a reduction in the starting dose of capecitabine/5-FU can be considered to avoid severe toxicity. There is no recommendation for a specific starting dose or maintenance dose in people with partial DPD activity. It has been reported that the variants *DPYD*2A* and *DPYD*13* lead to greater reduction in DPD enzymatic activity than any other *DPYD* variants and therefore have a much higher risk of having side effects.

Currently there is no evidence that states that a reduced capecitabine/5-FU dose is associated with lower drug toxicity/better drug efficacy in patients with partial DPD deficiency. Capecitabine/5-FU dose can be only increased in the absence of serious toxicity with careful monitoring of the patient. It might therefore take longer to recover from the side effects of chemotherapy if you have low or complete absence of DPD activity.

Our analysis identified, rs114045355 (1:97464825) as a *cis*-eQTL for *DPYD*. This implies that the variant allele T of the *cis*-eQTL is estimated to increase the expression of *DPYD* significantly. Therefore, if DPD activity is completely absent, and if the individual has the variant allele T of rs114045355, it is hypothesised that capecitabine/5-FU must not be administered. Since all transcripts of *DPYD* have zero DPD activity, the presence of the *cis*-eQTL will only increase the extent of DPD deficiency in that individual. Therefore, in such individuals, it is suggested that an alternate drug must be recommended. In people with DPD deficiency, it is postulated that the *cis*-eQTL will increase the expression of *DPYD* transcripts with deficient DPD significantly. This, in turn, might improve the clearance of the drugs from circulation and lower the probability of developing ADRs. In such an instance, the starting dose of capecitabine/5-FU might need to be reduced to further lower drug toxicity and development of ADRs. However, in people with normal DPD activity, as the *cis*-eQTL is predicted to increase the production of DPD, it is assumed that 5-FU may be degraded/cleared out a lot faster. Therefore, it is advised that the dose of capecitabine and 5-FU must be increased in people with normal DPD activity and rs114045355T, in order to achieve optimal drug efficacy levels. Though changes to capecitabine and 5-FU drug dosing have been suggested with respect to the presence of the variant allele of the *cis*-eQTL, these suggestions are made purely based on bioinformatic predictions and therefore must be clinically and functionally validated.

4.1.2.2 Variant annotations for DPYD and the effect of the eQTL

The variant annotations for *DPYD* are available at https://www.pharmgkb.org/gene/PA145/variantAnnotation. In patients with colorectal cancer, rs12132152A and rs76387818A are associated with increased risk of experiencing

drug toxicity like hand-foot syndrome, when treated with capecitabine as compared to patients with the G allele (Rosmarin *et al.*, 2015). However, in patients with colorectal cancer, rs12022243C decreases the likelihood of experiencing drug toxicity like diarrhoea, when treated with capecitabine as compared to patients with the T allele (Rosmarin *et al.*, 2015). Gastric and colon carcinoma patients with rs1801159C, when treated with fluoropyrimidine-based chemotherapy, have an increased likelihood of developing drug toxicity, and decreased clearance of the drug as compared to patients with the T allele (Zhang *et al.*, 2007). Fluoropyrimidines are often used in combination chemotherapy regimens like the FOLFOX (fluorouracil, leucovorin and oxaliplatin) or with other drugs such as bevacizumab, cetuximab, and raltitrexed. Cancer patients with *DPYD*6* allele, when treated with fluoropyrimidine-based chemotherapy, have been found to have decreased metabolism of fluorouracil which results in an increased risk for drug toxicities as compared to patients with the *DPYD*1* allele (Kuilenburg *et al.*, 2016; Thomas *et al.*, 2016). Cancer patients with *DPYD*2A* allele, when treated with fluoropyrimidine-based chemotherapy, have DPYD deficiency which results in decreased clearance of fluoropyrimidine drugs. They have been found to be at risk for severe or fatal drug toxicity as compared to patients with the *DPYD*1* allele (Deenen *et al.*, 2011; SAIF, 2013). Cancer patients with rs67376798A treated with fluoropyrimidine-based chemotherapy have been found to have complete DPYD deficiency and decreased clearance of fluoropyrimidine drugs which will lead to an increased risk severity of drug toxicity as compared to patients with the T allele (SAIF, 2013; Terrazzino *et al.*, 2013). Cancer patients with the *DPYD*13* allele when treated with fluoropyrimidine-based chemotherapy may have DYPD deficiency which will lead to increased risk and severity of drug toxicity as compared to patients with the *DPYD*1* allele (Johnson *et al.*, 2002; Morel *et al.*, 2006; Meulendijks *et al.*, 2015). Patients with the *DPYD*3* allele may have decreased DPYD activity as compared to those with the *DPYD*1* allele (Offer *et al.*, 2014). Patients with DPYD*7 allele are associated with decreased DPYD activity as compared to those with the *DPYD*1* allele (Offer *et al.*, 2014). *DPYD*10* allele (rs1801268) and *DPYD*8* allele (rs1801266) are associated with decreased DPYD activity as compared to those with the allele *DPYD*1* (Offer *et al.*, 2014). Cancer patients with rs75017182C when treated with fluoropyrimidine-based chemotherapy have shown a high risk of drug toxicity and increased likelihood of dose intervention as compared to patients with the G allele may be required (Jennings *et al.*, 2013; Froehlich *et al.*, 2015).

Similarly, cancer patients with rs78060119A are at increased risk for drug toxicity when receiving fluorouracil compared to patients with the C allele (Offer *et al.*, 2014).

From the examples listed above, it can be inferred that cancer patients with *DPYD*2A*, *DPYD*3*, *DPYD*6*, *DPYD*7*, *DPYD*8,* and *DPYD*13* alleles are associated with decreased DPYD activity and therefore when treated with fluoropyrimidine-based chemotherapy, because of DYPD deficiency, have increased risk of drug toxicity as compared to patients with the *DPYD*1* allele. Since the *cis*-eQTL, rs114045355T is hypothesised to significantly increase the expression of *DPYD*, in patients with completely absent DPD activity and rs114045355T, it is suggested that fluoropyrimidine-based chemotherapy be avoided. Another treatment regime must be advised to avoid the occurrence of severe ADRs. In patients with partial DPD activity and rs114045355T, fluoropyrimidine-based chemotherapy may be administered with extreme caution as the increased expression of *DPYD* will improve the metabolism of the drugs and lower the risk of developing ADRs.

If cancer patients with alleles that confer an increased risk for drug toxicity when receiving fluoropyrimidine based treatment (for example: rs75017182, rs78060119, rs12132152, rs76387818, rs1801159) also have the *cis*-eQTL, the drug dosage levels may need to be lowered so that the occurrence of ADRs may be monitored and controlled. Since the *cis*-eQTL is predicted to increase the expression of the *DPYD* transcript with the risk allele, extreme caution must be exercised when administering the drug. If the onset of ADRs is sudden despite low drug dose, it is suggested that treatment be stopped immediately.

On the other hand, if a cancer patient has an allele that lowers the risk of developing ADRs like rs12022243 and the *cis*-eQTL, it is hypothesised that the likelihood of them developing ADRs is lowered significantly. This is so because the *cis*-eQTL will significantly increase the expression of the *DPYD*.

All suggestions made with respect to modifications in drug dose must be clinically and functionally validated as they have been made based on bioinformatic predictions.

4.1.3 UGT1A1

UGT1A1 is an ADME gene located at 2q37 that codes for a phase II drug metabolising enzyme, uridine diphosphate glucuronosyl transferase1A1. It plays a major role in the conjugation and subsequent elimination of potentially toxic xenobiotics, carcinogens and pharmaceutical drugs (Schulz *et al.*, 2009; Barbarino *et al.*, 2014). Different alleles of *UGT1A1* confer different enzyme activity and depending on the genotype of *UGT1A1*, individuals are classified as extensive metaboliser, intermediate metaboliser or poor metaboliser of drugs (Gammal *et al.*, 2016).

rs28967009 (2:234515276) was identified as a *cis*-eQTL for *UGT1A1* in this analysis with a p-value of 5.73E-07, FDR of 0.006 and a beta-value of 1328. This implies that the variant allele A of the *cis*-eQTL significantly increases the expression of *UGT1A1* (1328 times on average). The effect of this *cis*-eQTL on *UGT1A1* is illustrated in Figure 4.3. The effect of the *cis*-eQTL on the metabolism of the antiretroviral dolutegravir and the chemotherapeutic irinotecan is discussed in detail in the sections below.

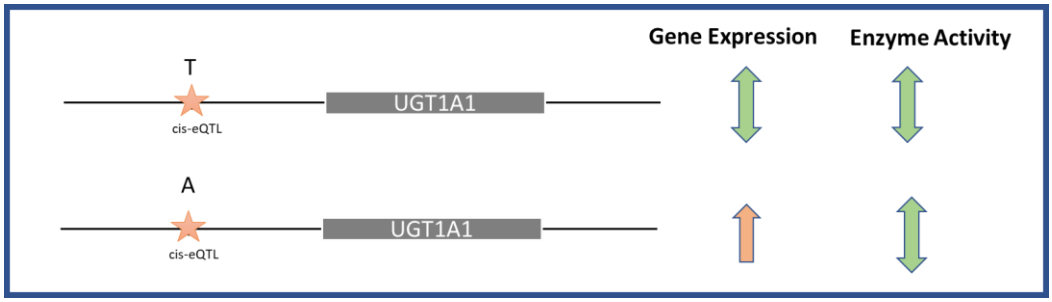


Figure 4.3: The effect of the variant allele A of the *cis*-eQTL, rs28967009, on UGT1A1 enzyme activity. The reference allele, T, is associated with normal (usual) *UGT1A1* expression whereas the variant allele, A, is associated with increased expression of *UGT1A1*.

4.1.3.1 Dolutegravir and UGT1A1

Dolutegravir is an anti-retroviral drug that is indicated for the treatment of HIV-1 infection, in combination with other antiretroviral agents (Yagura *et al.*, 2017). It belongs to a class of antiretroviral agents called integrase inhibitors. When human immunodeficiency virus (HIV) infects a cell, it combines its viral genetic code with the human genetic code through a process called integration. It uses the enzyme integrase for this purpose. Dolutegravir works by blocking integrase. When integrase is blocked, HIV cannot make more copies of itself. Dolutegravir is primarily metabolised in the liver by UGT1A1 to form dolutegravir glucuronide, which is the excretable form of the drug (Castellino *et al.*, 2013; Yagura *et al.*, 2017). Dolutegravir is widely used in SA and other developing countries in combination with tenofovir and lamivudine as first-line ART (Walmsley *et al.*, 2013; Moorhouse *et al.*, 2018; Venter *et al.*, 2019). It is highly effective in reducing the HIV viral load thereby lowering the chance of HIV transmission and developing AIDS and HIV-related illnesses or complications like infections and cancer. Common ADRs associated with dolutegravir are neuropsychiatric adverse events (NPAEs) like mood changes, depression and anxiety and gastrointestinal adverse events like nausea and vomiting (Elzi *et al.*, 2017; Lepik *et al.*, 2018; Cuzin *et al.*, 2019; Llibre *et al.*, 2019).

People with *UGT1A1* genotypes that confer reduced UGT1A1 activity have been shown to have 32% lower clearance and 46% higher AUC when compared to patients with *UGT1A1* genotypes that relate to normal UGT1A1 activity (Mercadel *et al.*, 2014). In simpler terms, individuals with *UGT1A1* genotypes that result in poor metabolism of the drug have decreased clearance and increased AUC as compared to those who have *UGT1A1* genotypes that confer normal dolutegravir metabolism.

The presence of the *cis*-eQTL rs28967009A is predicted to lower the occurrence of ADRs in UGT1A1 poor and intermediate metabolisers because increased production of UGT1A1 enzyme will help in faster drug clearance. They will metabolise dolutegravir faster than UGT1A1 intermediate and poor metabolisers without the *cis*-eQTL. However, in UGT1A1 extensive metabolisers with rs28967009A, dolutegravir dose may need to be increased, as

the *cis*-eQTL might result in increased/faster glucuronidation of the drug due to increased *UGT1A1* expression. In such individuals the drug might be cleared from the system before it exerts the desired effect on the body and therefore dose modification might be needed. Changes to dolutegravir drug dosing have been made with respect to the presence of the variant allele of the *cis*-eQTL purely based on bioinformatic predictions and must be clinically and functionally validated.

4.1.3.2 Irinotecan and UGT1A1

Irinotecan is an anti-cancer drug that is widely prescribed along with other drugs to treat advanced/metastatic and/or recurrent colorectal cancer (Ramesh *et al.*, 2010). It is a topoisomerase I inhibitor. The topoisomerase I enzyme is a nuclear enzyme that catalyses nuclear processes like DNA replication, DNA repair, recombination, and regulation of DNA supercoiling. DNA topoisomerase I enzyme relieves torsional strain in DNA by inducing reversible single strand breaks to the DNA helix backbone. This allows single DNA strands to pass through the break. The 3' terminus of the broken DNA strand then binds covalently with topoisomerase I enzyme to form a cleavable complex which is a catalytic intermediate. Topoisomerase then reattaches the broken DNA strands to form chemically unaltered topoisomers that allows for transcription to proceed (Dean, 2018).

UGT1A1 is involved in the glucuronidation of SN-38, the active form of irinotecan. SN-38 is 100-1000 times more cytotoxic than irinotecan (Mathijssen *et al.*, 2001). SN-38 prevents the reattachment of single strand breaks by binding to the cleavable complex to form what is known as a drug-topoisomerase-DNA complex. This halts the DNA replication fork from moving and arrests DNA replication by introducing irreparable DNA damage and cell apoptosis (Dean, 2018). Glucuronidation of SN-38 results in the formation of SN-38 glucuronide, the excretable form of the drug (Dias *et al.*, 2012).

Irinotecan treatment is generally limited by the high incidence of drug-toxicity due to high levels of SN-38 but would not be effective if the circulation levels were below the therapeutic

range. The common ADRs of irinotecan, severe neutropenia, fever, asthenia and diarrhoea, are caused by prolonged exposure to SN-38. These side effects can be severe enough to lead to the prescription of a reduced drug dose or to discontinuance of drug use (Lankisch *et al.*, 2008; Barbarino *et al.*, 2014). The rate of severe ADRs associated with irinotecan treatment is around 20-25% (Tam *et al.*, 2009). Some patients develop fever and severe neutropenia following treatment. Of those patients, around 7% die from these complications (Obradovic *et al.*, 2008; Liu *et al.*, 2014).

Genetic variants of *UGT1A1* associated with decreased UGT enzymatic activity increases the risk of irinotecan toxicity. *UGT1A1*28* is one such variant that results in reduced excretion of irinotecan. It is commonly found in African-Americans with an allele frequency of 0.42–0.45 (Dean, 2018). A 35% reduction in transcriptional activity has been identified when a single copy of the **28* allele is present. This doubles to 70% reduction when 2 copies (**28/*28*, homozygous) are present (Barbarino *et al.*, 2014). This leads to irinotecan toxicity because of increased active irinotecan metabolites in the blood. *UGT1A1*36* (associated with increased enzyme activity) and *UGT1A1*37* (associated with reduced enzyme activity) alleles occur at estimated allele frequencies of 0.03–0.10 and 0.02–0.07, respectively. These 2 alleles are almost exclusively present in populations of African origin(Dean, 2018). *UGT1A1*6* is another variant predominantly found in Asian populations that results in reduced UGT1A1 enzymatic activity.

The drug label for irinotecan (CAMPTOSAR, indicated for the treatment of metastatic colon or rectum carcinoma) by the US FDA, HCSC, PMDA, Swissmedic and EMA indicates that when administered as a combination therapy or as a single-agent, the starting dose of CAMPTOSAR must be reduced by at least one level in patients who have two copies of the *UGT1A1*28* allele and in patients who have experienced hematologic toxicity with previous treatment. The refined dose is made keeping the individual's tolerance to treatment in consideration. Individuals with the *UGT1A1*28/*28* genotype are at increased risk for neutropenia. Blood count and liver function tests in such individuals must be monitored closely. Irinotecan is commonly prescribed in South Africa (SA) to treat colorectal cancer (CRC) as it is the fourth

most common and sixth most lethal cancer in SA (Brand *et al.*, 2018; Herbst *et al.*, 2020). The disease burden of CRC in SA is high.

The drug label for irinotecan (ONIVYDE, indicated for the treatment of metastatic pancreatic adenocarcinoma) by the US FDA, HCSC, PMDA, Swissmedic and EMA states that the recommended starting dose of ONIVYDE in patients homozygous for the *UGT1A1*28* allele and in patients who have experienced hematologic toxicity with previous treatment, must be reduced to 50mg/m^2 administered by IV infusion over 90 minutes. The dose can be increased to a maximum of 70mg/m^2 in subsequent cycles based on the tolerance levels of the patient.

Since the variant allele A of the *cis*-eQTL rs28967009, significantly increases the expression of *UGT1A1*, in UGT1A1 extensive and ultra-rapid metabolisers, SN-38 is converted to SN-38 glucuronide before it reaches therapeutic levels. Therefore, in *UGT1A1* ultra-rapid and extensive metabolisers with rs28967009A, irinotecan dose might need to be increased, as the *cis*-eQTL might culminate in increased glucuronidation of the drug due to increased expression of *UGT1A1*. In such individuals the drug might get excreted before it gets to exert its effect and therefore if the drug dose is not increased, drug efficacy levels will be poor. However, UGT1A1 poor and intermediate metabolisers with the *cis*-eQTL rs28967009A might show improved enzyme activity because of increased *UGT1A1* expression. Irinotecan might get glucuronidated a lot faster than without the *cis*-eQTL thereby lowering the chance of ADR occurrence. Although suggestions have been made to alter the dosing of irinotecan in individuals with the variant allele of the *cis*-eQTL, they are solely based on bioinformatic predictions and must be clinically validated.

4.1.3.3 Variant annotations for UGT1A1 and the effect of the eQTL

https://www.pharmgkb.org/gene/PA420/variantAnnotation lists the variant annotations for *UGT1A1* in PharmGKB. Atazanavir and ritonavir are used as combination ART for the treatment of HIV in South Africa (Achenbach *et al.*, 2011). Atazanavir inhibits UGT1A1, preventing bilirubin to be eliminated in patients, eventually leading to hyperbilirubinemia and

jaundice. When this occurs, atazanavir treatment must be discontinued. *UGT1A1* PMs are at the highest risk for bilirubin-related discontinuation of atazanavir (Lankisch *et al.*, 2008; Turatti *et al.*, 2011; Gammal *et al.*, 2016). Variants in *UGT1A1* are associated with increased risk of nephrolithiasis (rs8330G, rs1042640G, rs10929303T), increased likelihood of developing hyperbilirubinemia (rs887829T), increased chance of discontinuation of atazanavir treatment (rs887829T) and increased metabolism of atorvastatin (rs887829T), a drug used in the treatment cardiovascular disease (Johnson *et al.*, 2014; Nishijima *et al.*, 2014; Vardhanabhuti *et al.*, 2015; Turner *et al.*, 2020). rs4148323A is associated with decreased metabolism of SN-38 in people with neoplasms, decreased overall survival (OS) and progression free survival (PFS) (Han *et al.*, 2006; Jada *et al.*, 2007; Takano *et al.*, 2009). *UGT1A1*28* is associated with slower elimination of raloxifene (prescribed to reduce the risk of invasive breast cancer) and development of hyperbilirubinemia when treated with tranilast, an antiallergic medication (Danoff *et al.*, 2004; Palomaki *et al.*, 2009; Trontelj *et al.*, 2009). *UGT1A1*6* allele is associated with lower clearance of etoposide (an anticancer agent) in people of African ancestry (Kishi *et al.*, 2004). Since the *cis*-eQTL rs28967009A, significantly increases *UGT1A1* expression, its presence along with existing variants of *UGT1A1* that increase susceptibility to adverse drug reactions, might lower the risk of developing these adverse drug reactions. For example, as rs887829T is associated with increased metabolism of atorvastatin, an alternate drug might need to be given in people who have both rs887829T and the A allele of the *cis*-eQTL as the drug will be cleared much faster due to increased *UGT1A1* expression. These recommendations have been made purely based on bioinformatic predictions and therefore must be clinically and functionally validated.

4.1.4 CYP3A5

CYP3A5 is part of a cluster of cytochrome P450 genes on chromosome 7q21.1 and encodes a monooxygenase that metabolises drugs as well as the steroid hormones testosterone and progesterone. *CYP3A5* expression and enzymatic activity are modulated by polymorphisms in the gene. Depending on the genotype of *CYP3A5*, individuals are classified as extensive metaboliser, intermediate metaboliser or poor metaboliser of drugs (Birdwell *et al.*, 2015).

This analysis identified rs143210517 (7:100085809) as a *cis*-eQTL for *CYP3A5* with a p-value of 3.76E-08, FDR of 0.001 and a beta-value of 16158 implying that the variant allele T of the *cis*-eQTL will significantly increase the expression of *CYP3A5* (16158 times on average). The effect of this *cis*-eQTL on *CYP3A5* is illustrated in Figure 4.4. The effect of the *cis*-eQTL on the metabolism of the antiretroviral maraviroc is discussed in detail below.

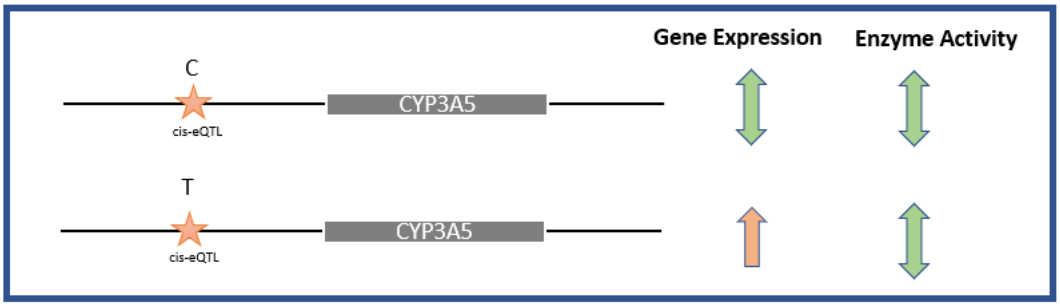


Figure 4.4: The effect of the variant allele T of the *cis*-eQTL, rs143210517, on CYP3A5 enzyme activity. The reference allele, C, is associated with normal (usual) *CYP3A5* expression whereas the variant allele, T, is associated with increased expression of *CYP3A5*.

4.1.4.1 Maraviroc and CYP3A5

Maraviroc is a U.S. Food and Drug Administration (FDA) approved drug for the treatment of human immunodeficiency virus (HIV). Maraviroc is usually administered in combination with other anti-HIV drugs. It exerts its effect by inhibiting the entry of the virus into human host cells by blocking the interaction between a viral envelope glycoprotein called gp120 and a coreceptor that HIV uses to enter human host cells called chemokine C-C motif receptor 5 (CCR5) (Dorr *et al.*, 2005). CCR5 is predominantly used as a coreceptor by HIV strains during transmission and early stages of the disease. Maraviroc is the only available CCR5 antagonist in the market. Since maraviroc penetrates the primary sites of sexual transmission (male and female reproductive tracts and colorectal tissues), this anti-retroviral is also considered as a potential treatment to prevent the onset of HIV (Brown *et al.*, 2011).

Maraviroc is orally administered and undergoes extensive metabolism via *N*-dealkylation and oxygenation. *N*-dealkylation is primarily mediated by *CYP3A4* and oxygenation is mediated by

CYP3A5. As oxidative metabolism plays an important role in the elimination of maraviroc from the body, polymorphisms in *CYP3A5* determine how long maraviroc remains in circulation in the body (Lu *et al.*, 2014). *CYP3A5*1* is the wild-type allele and is associated with the highest CYP3A5 activity/protein expression (Birdwell *et al.*, 2015). Other *CYP3A5* alleles like *CYP3A5*3*, *5, **6,* and **7* result in no CYP3A5 activity/protein expression (Birdwell *et al.*, 2015). *CYP3A5 *8* and **9* result in reduced CYP3A5 enzyme activity (Birdwell *et al.*, 2015). Due to differences in allele frequency distribution, CYP3A5 expression varies greatly in different populations. 80-90% of European Americans are CYP3A5 non-expressors. *CYP3A5*3* is very common in this ethnicity (Daly, 2006). Majority of African-Americans are CYP3A5 expressors. They carry at least one *CYP3A5*1* allele and approximately 45% of the population are homozygous *CYP3A5*1* carriers (Xie *et al.*, 2004). *CYP3A5* genotype contributes majorly to the variability of maraviroc exposure. *CYP3A5*1* homozygous carriers have the least exposure to maraviroc followed by heterozygous carriers of the dysfunctional/non-functional allele followed by homozygous carriers of the dysfunctional/non-functional allele.

Since more African-Americans/Africans are infected with HIV and because they are at a much higher risk of contracting the disease than people of other ethnicities, the fact that *CYP3A5*1* homozygous carriers are associated with reduced maraviroc exposure is quite a significant finding (Lu *et al.*, 2014). If the dosage of maraviroc is not increased in homozygous *CYP3A5*1* allele carriers, people with this genotype will be at risk of HIV treatment failure. Swissmedic mentions this in the drug annotation for maraviroc and PharmGKB lists this as an 'actionable PGx'.

It is suggested that CYP3A5 extensive metabolisers with the *cis*-eQTL rs143210517T, may need an increase in maraviroc dose, as the *cis*-eQTL may result in increased oxidation of the drug due to increased *CYP3A5* expression. In such individuals maraviroc might be cleared from the system before it reaches therapeutic levels and exerts the desired effect on the body. Therefore, CYP3A5 extensive metabolisers with rs143210517T may need maraviroc in much higher doses in order to attain desirable drug efficacy levels. For CYP3A5 poor metabolisers who are homozygous carriers of the non-functional alleles with the *cis*-eQTL rs143210517T,

an alternate treatment regime might need to be suggested since CYP3A5 is not produced at all. In CYP3A5 poor metabolisers with heterozygous dysfunctional/non-functional alleles and the *cis*-eQTL and in CYP3A5 intermediate metabolisers, although the *cis*-eQTL is predicted to improve enzyme activity due to increased *CYP3A5* expression, the recommended dosage levels of maraviroc might need to be reduced to lower the probability of ADR occurrence. Suggestions to changes in maraviroc dosing is made based on bioinformatic methods and therefore must be clinically and functionally validated.

4.1.4.2 Variant annotations for CYP3A5 and the effect of the eQTL

Variant annotations for *CYP3A5* can be accessed from PharmGKB at https://www.pharmgkb.org/gene/PA131/variantAnnotation. In HIV patients treated with maraviroc or atazanavir boosted with ritonavir, the wild-type *CYP3A5 *1* allele is associated with increased metabolism of both treatments. HIV patients with any combination of two non-functional *CYP3A5* alleles (**3*, **6*, or **7* alleles), will metabolise maraviroc and atazanavir-ritonavir combination much more slowly compared to individuals with just one *CYP3A5 *1* allele or the *CYP3A5 *1/*1* genotype (Anderson *et al.*, 2009; Savic *et al.*, 2012; Vourvahis *et al.*, 2019). *CYP3A5 *1* allele is also found to be associated with increased metabolism of atorvastatin (a drug used to treat cardiovascular disease) and amlodipine (a drug used to treat hypertension) when compared to the non-functional *CYP3A5* alleles (Wilke *et al.*, 2005; Willrich *et al.*, 2008, 2009; Saiz-Rodríguez *et al.*, 2020). In patients with HIV, cardiovascular disease or hypertension with the wild-type functional allele, *CYP3A5*1* and the *cis*-eQTL rs143210517, a significant increase in *CYP3A5* expression is predicted. Therefore, consequently these individuals may show increased metabolism of the drugs maraviroc, atazanavir boosted with ritonavir, atorvastatin and amlodipine. Since the CYP3A5 enzyme might be more active in these people due to the *cis*-eQTL, the dosage levels of the drugs might need to be increased to attain ideal drug efficacy levels.

Hypercholesterolemia patients treated with atorvastatin have a better response when they have the rs17161788C (Willrich *et al.*, 2008). If these patients have the *cis*-eQTL as well, it is predicted that the treatment response to atorvastatin will also significantly improve. This

implies that patients with hypercholesterolemia who have both the *cis*-eQTL rs143210517T and rs17161788C may have a much higher likelihood of good treatment response when treated with atorvastatin.

HIV patients with the *CYP3A5*3* allele are associated with increased metabolism of nevirapine and lopinavir, anti-retrovirals that are used for the treatment of HIV (Ciccacci *et al.*, 2009; Zhang *et al.*, 2013). The same allele is associated with decreased metabolism of efavirenz, which is yet another anti-retroviral (Haas *et al.*, 2004). Therefore, HIV patients with the *cis*-eQTL rs143210517T who are treated with nevirapine and lopinavir, might have significantly increased expression of *CYP3A5* due to the *cis*-eQTL, which may result in the drugs being metabolised faster. Hence the drug doses of nevirapine and lopinavir might have to be increased to attain ideal drug efficacy levels. However, in the HIV patients treated with efavirenz, who have the *cis*-eQTL rs143210517T, as the *cis*-eQTL significantly increases CYP3A5 enzyme expression, though maraviroc will be metabolised faster than in people without the *cis*-eQTL, the drug dose of efavirenz might have to be decreased to lower ADR occurrence.

Although recommendations have been to modify drug dose based on the presence of the variant allele of the *cis*-eQTL, they have been made solely based on bioinformatic predictions and must be clinically and functionally validated.

4.1.5 SULT1A1

SULT1A1 gene resides on chromosome 16 and has 7 coding exons and 2 alternatively spliced upstream non-coding exons. It encodes a 295 amino acid protein, the sulphotransferase 1A1 (SULT1A1), a phase II biotransformation enzyme (Jakoby *et al.*, 1990; Hildebrandt *et al.*, 2009). Sulphotransferase 1A1 is the most widely expressed member of the sulphotransferase family (Hildebrandt *et al.*, 2009). Polymorphisms in *SULT1A1* have the potential to influence protein function resulting in SULT1A1 enzyme isoforms with both increased enzyme activity and decreased enzyme activity.

rs1000598999 (16:27948796) was identified as a *cis*-eQTL for *SULT1A1* with a p-value of 5.33E-07, FDR of 0.005 and a beta-value of 47 implying that the variant allele T of the *cis*-eQTL is associated with significant increase in the expression of *SULT1A1* (47 times on average). The effect of this *cis*-eQTL on *SULT1A1* is illustrated in Figure 4.5. The effect of the *cis*-eQTL on the metabolism of commonly prescribed analgesics like acetaminophen, naproxen and tapentadol is discussed below.

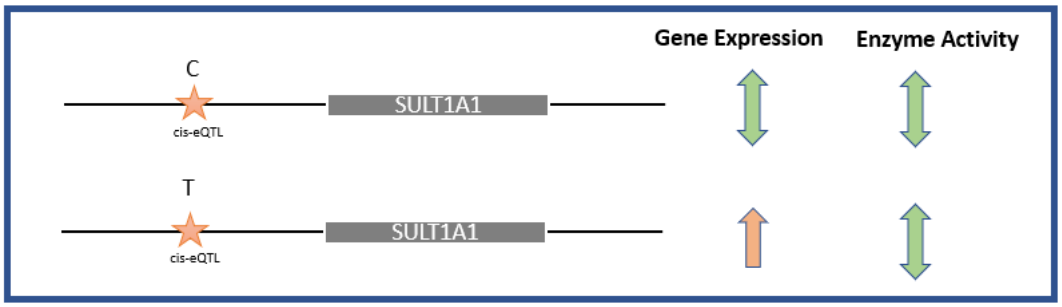


Figure 4.5: The effect of the variant allele T of the *cis*-eQTL, rs1000598999, on SULT1A1 enzyme activity. The reference allele, C, is associated with normal (usual) *SULT1A1* expression whereas the variant allele, T, is associated with increased expression of *SULT1A*

4.1.5.1 Acetaminophen, naproxen tapentadol and SULT1A1

Acetaminophen is one of the most prescribed medication for pain and fever. It is also one of the most purchased over-the-counter medications. Acetaminophen metabolism takes place predominantly in the liver where the majority of the drug is either glucuronidated or sulphated before it is excreted in the urine (McGill *et al.*, 2013; Yamamoto *et al.*, 2015). About 25–35% of the therapeutic dose of acetaminophen is recovered in its sulphated form and 50–70% is recovered in its glucuronidated form (McGill *et al.*, 2013). An overdose of acetaminophen causes severe liver toxicity (Davidson *et al.*, 1966; Larson *et al.*, 2005; Wei *et al.*, 2007).

Like acetaminophen, naproxen (desmethylnaproxen) and tapentadol are also quite commonly used as prescribed and over-the-counter medication for the management of pain

and fever across various age groups (Rasool *et al.*, 2019). Combinations of these drugs are used for the clinical management of acute and chronic pain. SULT1A1 is the primary enzyme responsible for the sulfation of acetaminophen, naproxen and tapentadol (Falany *et al.*, 2005; Yamamoto *et al.*, 2015; Bairam *et al.*, 2017). SNPs in *SULT1A1* result in significant and dramatic differences in their catalytic activity of its different substrates. Thus, these polymorphisms can affect how acetaminophen, naproxen and tapentadol are metabolised in the body.

As *SULT1A1* is involved in the sulfation of acetaminophen, naproxen and tapentadol, the presence of the variant allele T of the *cis*-eQTL rs1000598999 is expected to significantly increase the rate at which sulphation of these drugs occur because of increased *SULT1A1* expression. Therefore, in people with *SULT1A1* variants that confer increased enzymatic activity and the *cis*-eQTL rs1000598999T, these drugs may need to be administered in increased doses in order to achieve therapeutic levels as they might be cleared from the system faster due to the *cis*-eQTL. Despite increasing the dose if adequate drug efficacy is not achieved, alternate drugs for pain management might need to be suggested. However, in individuals with the *cis*-eQTL and *SULT1A1* variants that confer decreased SULT1A1 activity, the *cis*-eQTL rs1000598999T may lower the occurrence of ADRs associated with the drugs as they might be cleared faster from the system. A reduction of drug dose might further help reduce ADRs in these people. Although recommendations have been made to alter the doses of acetaminophen, naproxen and tapentadol, they have been made solely based on bioinformatic predictions and must be clinically and functionally validated.

4.1.5.2 Variant annotations for SULT1A1 and the effect of the eQTL

PharmGKB has variant annotations for *SULT1A1* which can be accessed at https://www.pharmgkb.org/gene/PA343/variantAnnotation. To cite a few examples, patients with A allele of rs767487725, a SNP in *SULT1A1*, is found to have increased sulfation of acetaminophen, naproxen and tapentadol as compared to patients with the C allele (Rasool *et al.*, 2019). Similarly, rs544820732G and rs72547527C are also found to have increased sulfation of acetaminophen, naproxen and tapentadol as compared to A allele and T allele respectively (Rasool *et al.*, 2019). rs28374453A and rs765399160C are found to be associated

with decreased metabolism of acetaminophen and naproxen compared to G and T allele respectively (Rasool *et al.*, 2019).

If an individual has rs767487725A or rs544820732G (both associated with increased sulfation of the drugs) and the *cis*-eQTL rs1000598999T, the sulfation of the drugs, acetaminophen, naproxen and tapentadol, may occur faster. Thus, in order to attain ideal drug efficacy levels in these patients, the drug doses of acetaminophen, naproxen and tapentadol might need to be increased. Despite increasing drug dose, if therapeutic levels are not reached, alternate pain medication may need to be suggested.

However, if an individual has rs28374453A or rs765399160C (associated with decreased metabolism of the drugs) and the *cis*-eQTL rs1000598999T, the sulfation of the drugs, acetaminophen and naproxen may occur faster than in individuals without the *cis*-eQTL. The *cis*-eQTL may make such people better tolerant to acetaminophen and naproxen. Thus, in order to further improve the metabolism of these drugs in such patients, the dosage levels of the drugs might need to be decreased. By doing so, the probability of ADR occurrence might also be reduced.

Although suggestions have been made to modify drug doses of acetaminophen, naproxen and tapentadol due to the presence of the variant allele of the *cis*-eQTL, they have to be clinically validated since they are based on bioinformatic predictions.

4.1.6 UGT2B17

Despite being listed as a core ADME gene, *UGT2B17* has no drug label annotations associated with it in PharmGKB. The gene also had no variant annotations in PharmGKB. Therefore, the effect of the *cis*-eQTL, rs917897719 (4:68462245) identified by this study, on *UGT2B17* despite being significant with p-value 1.46E-07, FDR 0.002 and beta value 3944 cannot be understood.

4.2 EXTENDED ADME GENES AND THEIR eQTLs

Of the 63 extended ADME genes for which *cis*-eQTLs were found only *UGT1A6*, *UGT2B15*, *ABCC3*, *ABCC4*, *XDH* and *CES1* were found to be involved in the metabolism of drugs listed in Tables 3.35-3.39. These drugs are commonly prescribed to combat the top ten causes of mortality in South Africa (SA) published by the Department of Statistics, South Africa, 2017. The top ten causes of mortality in SA include HIV and AIDS, tuberculosis, diabetes, hypertension and cardiovascular diseases. As analgesics are quite extensively used in SA, we also look at over-the-counter and prescription medicines.

4.2.1 UGT1A6

This study identified rs187641587 (2:235197314) as a *cis*-eQTL for *UGT1A6* with a significant p-value of 2.00E-21, beta value of 815 and a FDR of 2.67E-16 in imputed dataset A. In imputed dataset B, the p-value, beta value and FDR are 2.02E-20, 815 and 2.42E-15 respectively. This implies that the variant allele T of this *cis*-eQTL significantly increases the expression of *UGT1A6* (815 times on average). Table 3.35 lists out the variant annotations of *UGT1A6* and acetaminophen, an analgesic that is mostly bought over-the-counter. Individuals with the *UGT1A6 *2a* allele and rs187641587T, may have increased acetaminophen clearance than individuals who do not have the *cis*-eQTL. Thus, in order to attain ideal drug efficacy levels in individuals with the *UGT1A6 *2a* allele and the *cis*-eQTL for the gene, the drug dose levels of acetaminophen might have to be increased. If therapeutic levels of the drug and drug efficacy is still not attained despite increasing the drug dose, an alternate treatment should be recommended. Dosing recommendation made for acetaminophen in individuals with the variant allele of the *cis*-eQTL is solely based on bioinformatic computation and hence must be clinically and functionally validated.

4.2.2 ABCC3

rs146347791 (17:48566159) was identified as a *cis*-eQTL for *ABCC3* with a p-value of 2.54E-07, a FDR of 0.003 and a beta-value of 3,916 in imputed dataset A. It was also identified as a *cis*-eQTL in imputed dataset B with p-value of 4.44E-07, FDR of 0.005 and beta-value of 3,913.

In both datasets, the FDR and p-values are significant. This implies that the variant allele A of this *cis*-eQTL is associated with significant increase in the expression of *ABCC3* (3915 times on average). The variant annotations found for variants in *ABCC3* and the anti-cancer drugs cisplatin, cyclophosphamide, doxorubicin, methotrexate and vincristine are listed in Table 3.36. This implies that in children with neoplasms treated with cisplatin who have the G allele of rs1051640, the risk of ototoxicity increases compared to children who do not have the *cis*-eQTL. Similarly, in osteosarcoma patients with rs4148416T and the *cis*-eQTL, who are treated with cisplatin, cyclophosphamide, doxorubicin, methotrexate and vincristine, the risk of death is higher than in patients who do not have the *cis*-eQTL. Drug dosage levels might need to be decreased to minimise the risk of ototoxicity in cancer patients when treated with cisplatin or an alternative drug may need to be suggested. Similarly, to lower the risk of death when treated with cisplatin, cyclophosphamide, doxorubicin, methotrexate and vincristine in patients with osteosarcoma, either drug dosage levels need to be lowered, or an alternative treatment strategy must be suggested. Treatment modification in individuals with the variant allele of the *cis*-eQTL is recommended solely based on bioinformatic computation and hence must be clinically and functionally validated.

4.2.3 ABCC4

rs188869286 (13:95531528) was identified as a *cis*-eQTL for *ABCC4* with a p-value of 1.09E-11, a beta-value of 169 and a FDR of 4.79E-07 in imputed dataset A and in imputed dataset B with a p-value of 6.42E-12, a beta-value of 170 and a FDR of 2.96E-07. This implies that the variant allele G of this *cis*-eQTL is associated with significant increase in the expression of *ABCC4* (170 times on average). Annotations were found for *ABCC4* variants and anti-cancer drugs tenofovir, cyclophosphamide, doxorubicin, fluorouracil, cisplatin and irinotecan. They are listed in Table 3.37. This implies that patients treated with tenofovir who have the TT genotype of rs1751034 and the *cis*-eQTL might have higher tenofovir clearance compared to patients without the *cis*-eQTL. Therefore, in such patients, tenofovir dose may need to be increased to achieve ideal drug efficacy levels. Similarly in patients with rs1059751G and the *cis*-eQTL rs188869286G, the likelihood of developing kidney disease when treated with tenofovir is may be higher compared to patients without the *cis*-eQTL. In such patients, it is suggested that either the dosage of tenofovir must be lowered (to reduce the likelihood of

developing kidney disease) or an alternate treatment regime must be recommended. Patients with rs3742106C and the *cis*-eQTL may have increased tenofovir concentrations compared to patients with the AC or CC genotypes who do not have the *cis*-eQTL. The risk of developing ADRs like neutropenia and gastrointestinal toxicity in breast neoplasm patients with rs9561778T and the *cis*-eQTL, and are treated with the combination therapy, cyclophosphamide, doxorubicin and fluorouracil, or with just cyclophosphamide, may be much higher than in patients without the *cis*-eQTL. For such patients, dose adjustments to the drug therapy might need to be made to reduce or prevent the occurrence of the ADRs or alternative treatment strategies must be administered. Small cell carcinoma patients who are treated with cisplatin and irinotecan who have rs16950650T and the *cis*-eQTL rs188869286G may have lower overall survival than patients without the *cis*-eQTL. In these patients, alternate treatment regimens might need to be opted for better overall survival of patients with small cell carcinoma. The likelihood of developing kidney disease in HIV patients with rs1059751G and variant allele G of the *cis*-eQTL, who are treated with tenofovir, may be higher than in patients without the *cis*-eQTL. In such patients as well, a different treatment strategy must be suggested to lower the likelihood of developing kidney disease. Dosing recommendations in individuals with the variant allele of the *cis*-eQTL is based on bioinformatic calculations and therefore must be clinically and functionally validated.

4.2.4 XDH

Our study identified a *cis*-eQTL, rs557139987 (2:30688726) for *XDH* in imputed dataset A with a p-value of 1.46E-08, beta-value of 1,543 and FDR of 0.0003. A p-value of 2.61E-08, beta-value of 1,540 and FDR of 0.0005 was identified for the same *cis*-eQTL for *XDH* in imputed dataset B. This means that the variant allele C of the *cis*-eQTL is associated with significant increase in the expression of *XDH* (1542 times on average). Variant annotations found for XDH and didanosine, a drug used for the treatment of HIV infections are listed in Table 3.38.

The risk of developing noncirrhotic portal hypertension in HIV patients treated with didanosine who have genotype AA (rs1429376) or genotype AA (rs1594160) and the *cis*-eQTL rs557139987C, may be much higher compared to patients without the *cis*-eQTL. Thus, in both these type of HIV patients treated with didanosine, either the drug dosage levels might need

to be lowered to reduce the likelihood of developing noncirrhotic portal hypertension, or an alternate treatment regime must be suggested. Modification to treatment in individuals with the variant allele of the *cis*-eQTL is based on bioinformatic calculations and must hence be clinically and functionally validated

4.2.5 CES1

Table 3.39 lists out variant annotations found for *CES1* and capecitabine, a drug used to treat cancer. This study identified rs775020102 (16:55140242) as a *cis*-eQTL for *CES1* in both imputed dataset A and B. In imputed dataset A, for the p-value, beta-value and FDR for the *cis*-eQTL and CES1 are 4.00E-09, 5544 and 9.30E-05 respectively. For imputed dataset B, the p-value, beta-value and FDR are 1.23E-08, 5516 and 0.0002 respectively. This implies that the variant allele T of the *cis*-eQTL rs775020102, is associated with a significant increase in the expression of *CES1* (5530 times on average). Therefore, the risk of developing drug toxicity in patients with neoplasms who have the allele G (for rs2244613, rs2244614 and rs3217164) and the *cis*-eQTL, rs775020102T may be higher than in patients with neoplasm without the *cis*-eQTL. Therefore, in such patients capecitabine might need to be administered in lower doses to reduce the risk of developing drug toxicity or a different treatment regime must be suggested. Although treatment modification in individuals with the variant allele of the *cis*-eQTL is suggested, since it is purely based on bioinformatic predictions, it must be clinically and functionally validated.

4.3 SAME eQTL FOR DIFFERENT GENES (PLEIOTROPIC EFFECTS)

It is notable that in one instance the same *cis*-eQTL had an effect on gene expression of 5 genes (Uridine 5'-diphospho-glucuronosyltransferases, UGTs) and in another instance an eQTL affected gene expression of two different genes (Transporter associated with antigen processing 1 and 2, TAPs). This could be because the regulation of expression of genes of the same family are coordinated by a single variant. rs187641587 was found to be a *cis*-eQTL for *UGT1A6, UGT1A7, UGT1A8, UGT1A9* and *UGT1A10* (

Table 4.1). rs187641587 was found to increase the expression of *UGT1A6* the most, followed by *UGT1A7, UGT1A10, UGT1A9* and then *UGT1A8* (decreasing order of beta-value). Similarly, rs530350631 was found to be a *cis*-eQTL for *TAP1* and *TAP2* (Table 4.2). The expression levels

of *TAP1* were increased almost ten-folds in comparison with *TAP2* by the eQTL rs530350631. Thus, it seems that the regulation by an eQTL has a different effect size depending on the gene.

Table 4.1: Comparison of beta value, p value and FDR for rs187641587 (2:235197314) and *UGT1A6*, *UGT1A7*, *UGT1A8*, *UGT1A9* and *UGT1A10*

Chr coordinate	rsID	Gene	Beta value	p value	FDR
2:235197314	rs187641587	*UGT1A6*	815.66	2.00E-21	2.67E-16
2:235197314	rs187641587	*UGT1A7*	449.13	3.81E-13	2.25E-08
2:235197314	rs187641587	*UGT1A8*	84.30	3.92E-12	1.87E-07
2:235197314	rs187641587	*UGT1A9*	183.07	3.14E-11	1.26E-06
2:235197314	rs187641587	*UGT1A10*	421.49	1.97E-15	1.48E-10

Table 4.2: Comparison of beta value, p value and FDR for rs530350631 (6:32844125) and *TAP1* and *TAP2*

Chr coordinate	rsID	Gene	Beta value	p value	FDR
6:32844125	rs530350631	*TAP1*	3632.71	1.91E-12	9.67E-08
6:32844125	rs530350631	*TAP2*	318.86	8.92E-11	3.32E-06

4.4 ALLELE FREQUENCIES OF THE *cis*-eQTLs IN AFRICAN POPULATIONS

Allele frequencies of the identified *cis*-eQTLs for the core and extended ADME genes were obtained from African populations in the 1000 Genomes Project and the African Genome Variation Project (n=320). Of the six core ADME genes for which *cis*-eQTLs were identified (*CYP2C19, CYP3A5, DPYD, SULT1A1, UGT1A1,* and *UGT2B17*) only the *cis*-eQTLs identified for *UGT1A1* (Table 3.14) and *DPYD* (Table 3.15) were observed at appreciable frequencies in African populations. Of the 63 extended ADME genes for which *cis*-eQTLs were identified in this study, only *cis*-eQTLs for twelve extended ADME genes (*SOD2, CYP26A1, SLC13A3*, *ABCC3*, *SLC22A17, SLC7A5*, *ALDH1A2*, *CHST9*, *SLCO1A2, ABCC11, ABCB8*, *CBR3*) (Tables 3.16 - 3.27) were observed in various African populations.

The *cis*-eQTLs for *ALDH1A2, CHST9, SLCO1A2, ABCC11, ABCB8* and *CBR3* are very rare and only found in selected African populations. *Cis*-eQTLs for *ABCC3*, *SLC22A17*, *SLC7A5* and *ALDH1A2* are rare but are found in more African populations than the *cis*-eQTLs for *ALDH1A2, CHST9, SLCO1A2, ABCC11, ABCB8* and *CBR3*. The *cis*-eQTLs of *SOD2*, *CYP26A1* and *SLC13A3* are found in all the African populations that we looked at. However, the *cis*-eQTLs for *SOD2* and *CYP26A1* are more common in African populations than that of SLC*13A3*.

The allele frequencies of eQTLs are important as they would reflect the proportions of the populations likely to be affected by the up or down regulation of gene expression for selected ADME genes. The quantitative difference in the amount of mRNA and likely the corresponding protein, could have a profound effect on the metabolism, transport, and excretion of drugs. The identification and characterisation of genetic variation in regulatory regions of the genome are therefore important to better understand drug efficacy and potential toxicity and other adverse events. This area of pharmacogenomics is less well developed and there is especially very little data from African populations.

4.5 ALLELE FREQUENCIES OF THE *cis*-eQTLs IN DIFFERENT POPULATIONS

Table 3.28 shows a comparison of the allele frequencies of the *cis*-eQTLs found in African populations with the other four major population groups in the 1000 Genomes Project namely American, East Asian, EUR and South Asian. *Cis*-eQTLs for *SOD2* (0.28 in Africans), *CYP26A1* (0.20 in Africans) are found to be present at a higher frequency in the aggregated African group compared to the other four aggregate populations. The *cis*-eQTL for *UGT1A1* is found to be most common in Americans (0.07) followed by Africans (0.05), South Asians (0.04), East Asians (0.03) and Europeans (0.01). The *cis*-eQTL for *SLC13A3* is found only in Africans. The *cis*-eQTL for *DPYD* is not found in Africans but is found in Europeans and South Asians. From the comparison in Table 3.28, it can be inferred that the eQTLs exist in varying frequencies across different populations.

4.6 HOW IS THIS STUDY DIFFERENT FROM SCHRÖDER *et al* (2013)

Schröder *et al.*, 2013 identified *cis*- and *trans*-eQTLs from 149 Causcasian liver samples. Either the identified eQTL had to be located in an ADME gene, or the gene whose expression is modulated by the eQTL had to be an ADME gene or both had to be an ADME gene to be considered significant by Schröder *et al*.. The identified eQTLs were then subjected to a genome wide association analysis and the eQTLs identifed were compared to two previously published stidues, namely the Stuttgart and Seattle studies.

In this study, though the same dataset of 149 liver samples was used, a bioinformatic pipeline was built to generate a list of *cis*- and *trans*-eQTLs. This bioinformatic pipeline also does genotype and gene expression QC prior to generating the list of eQTLs from the dataset. Though the pipeline identifies both *cis*- and *trans*-eQTLs, as it is difficult to replicate *trans*-eQTLs across studies, this study focussed only on *cis*-eQTLs of core and extended ADME genes. Inorder to be identified as a *cis*- or *trans*-eQTL of an ADME gene, the gene who's expression the eQTLs modulate had to either be a core or an extended ADME gene. An in-depth insilico analysis was then done on the *cis*-eQTLs identified for any core or exteneded ADME genes. An important and novel feature of this study is that it looked at the allele frequencies of the identified eQTLs in Africans. It is then postulated how the identified *cis*-eQTLs can affect drug dosing regimes for drugs that are most commonly prescribed in Africa.

4.7 COMPARISON WITH SCHRÖDER ET AL. AND ZHONG ET AL.

When eQTLs for ADME genes identified by the pipeline were compared with eQTLs identified by Schröder *et al.* and Zhong *et al.*, Schröder *et al.* identified eQTLs for *CYP3A5*, *UGT2B17*, *ABCC11* and *FMO2* like the pipeline, albeit different ones. Schröder *et al.* identified two eQTLs for *CYP3A5* (rs1859690 and rs10242455), rs2070959 as an eQTL for *UGT2B17*, rs8056100 as an eQTL for *ABCC11* and two eQTLs for *FMO2* (rs1795240 and rs1736565). The pipeline identified rs143210517 as an eQTL for *CYP3A5*, rs917897719 for *UGT1A1*, rs74874632 for *ABCC11* and rs188999033 for *FMO2*. Different SNPs were identified as eQTLs for the genes probably because of different criteria used for QC, different number of samples removed from the study because of different QC criteria and different reference panels used for imputation.

Using imputed dataset A, eQTLs were identified for 62 of the same genes for which eQTLs were also identified by Schröder *et al.* eQTLs were identified for 49 of the same genes using imputed dataset B and Schröder *et al*. This shows that as the number of samples removed increases, the number of common eQTLs between the datasets decreases as expected. However the eQTLs identified for the genes were different probably because different imputation panels and QC methods were used in both studies. The genes for which eQTLs were identified by Schröder *et al* were then searched for the 297 ADME genes from PharmaADME. 13 ADME genes from PharmaADME were identified in the eQTLs identified by Schröder *et al*..

Using imputed dataset A, eQTLs were identified for the same 22 genes for which Zhong *et al.* also identified eQTLs using Matrix eQTL. Using imputed dataset A, eQTLs were identified for the same 22 genes for which Zhong *et al.* also identified eQTLs using LAMatrix. When comparing the genes for which eQTLs were identified using MatrixeQTL by Zhong *et al.* and imputed dataset B, eQTLs were identified for 17 common genes. Using imputed dataset B, eQTLs were identified for the same 18 genes for which Zhong *et al.* identified eQTLs using LAMatrix. eQTLs identified for the gene are however different. This might be because Zhong *et al.* used hepatocytes from African-Americans and the pipeline was run on Caucasian liver samples. Different QC steps and panels were also used for imputation.eQTLs were identified for only 6 ADME genes from PharmaADME by Zhong *et al.* using Matrix eQTL. However, eQTLs were identified for only 4 ADME genes by Zhong *et al* that were in PharmaADME.

4.8 REPRODUCIBILITY OF eQTLs

Though eQTLs have been generated for the same genes, reproducibility of eQTLs between the three studies is low. Low reproducibility of eQTLs between the three studies can be attributed to different QC methods, different genotyping and gene expression platforms employed, different covariates used, different reference panels used for imputation and different methods used to identify the eQTLs. The samples also probably contribute to low reproducibility because Zhong *et al.* used hepatocytes and Schroeder *et al.* used liver samples

excised during surgery to identify eQTLs. Low reproducibility of eQTLs between the three studies highlights the novelty of the eQTLs identified by this study. Therefore, the eQTLs identified by this study must be clinically and functionally validated.

4.9 RATIONALE FOR USING IMPUTED DATA

In this study, when the pipeline was run using the QCed datasets, dataset A and dataset B, no eQTLs were identified for core or extended ADME genes that were significant. Therefore, the decision was made to run the pipeline using the imputed datasets, imputed dataset A and imputed dataset B, and see whether any eQTLs could be identified for core and extended ADME genes. eQTLs were identified for both core and extended ADME genes that were significant. Imputed dataset A generated 6 *cis*-eQTLs for core ADME genes (Table 3.6) and 57 *cis*-eQTLs for extended ADME genes (Table 3.8). Imputed dataset B generated 6 *cis*-eQTLs for core ADME genes (Table 3.6) and 53 *cis*-eQTLs for extended ADME genes (Table 3.9). This clearly demonstrates the value imputation of genotype data has in identifying eQTLs.

4.10 USEFULNESS AND NOVELTY OF THE PIPELINE

A unique feature of the pipeline is that if corresponding genotype and microarray data for gene expression is available for any tissue, the pipeline can be used to generate eQTLs. The pipeline is available in the public domain. It is to be noted that the pipeline also does QC on the datasets, both genotype and gene expression data and generates *cis*- and *trans*-eQTLs. Since the pipeline uses Matrix eQTL to generate eQTLs, it is quite fast and efficient. From QC of both the genotype and gene expression datasets, generation of the input files needed to run Matrix eQTL, and identification of *cis*- and *trans*-eQTLs took around 20 minutes (19 minutes and 53 seconds to be precise).

4.11 ADME GENES AND THEIR EXPRESSION IN THE LIVER

Though *CYP24A1*, *GSTA3*, *UGT1A8* and *UGT1A10* were assumed to be expressed in the liver, they were found to not be expressed in the liver samples available in GTEx (Table 3.10). There was also no data available on the expression levels of *CYP11A1*, *GSTM5* and *SLC22A17* in GTEx (Table 3.10).

4.12 NUMBER OF eQTLs IDENTIFIED AND ITS CORRELATION TO SAMPLE SIZE.

The number of *cis*-eQTLs identified is assumed to be correlated to sample size. It is assumed that more the number the samples, more the number of *cis*-eQTLs identified. However, when the pipeline was run using 136 samples (dataset B), 36 *cis*-eQTLs (Table 3.2) were identified compared to the 33 *cis*-eQTLs identified from a larger sample size of 145 (dataset A) (Table 3.1). This is quite surprising and is not clearly understood.

4.13 COMMON eQTLs IDENTIFIED FROM DATASETS AND IMPUTED DATASETS

When an assessment was done to determine whether there were any common eQTLs between dataset A and imputed dataset A, only eight *cis*-eQTLs were found to be common between both datasets (figure 3.5). Similarly, only seven *cis*-eQTLs were found to be common between dataset B and imputed dataset B. The finding that all *cis*-eQTLs identified in datasets A and B are not present in imputed datasets A and B respectively is somewhat puzzling, as all eQTLs identified in dataset A and B are expected to be present in the post-imputation datasets, imputed dataset A and B respectively. A possible explanation is that the eQTLs not identified in the imputed datasets failed the significance cut off in the larger datasets.

4.14 VALIDATION OF THE IDENTIFIED eQTLs

Since the eQTLs identified in this study were not present in GTEx, QTLabse or in previously published studies done by Schroeder et al. and Zhong et al., it is important that these eQTLs be experimentally validated to support the bioinformatic predictions made by this study. Of the six *cis*-eQTLs identified for the core ADME genes, only two eQTLs, *cis*-eQTL for DPYD (rs114045355) and *cis*-eQTL for *UGT1A1* (rs28967009), were found to be associated with drug response related phenotypes using Oxford BIG Server.

4.15 REGULOMEDB AND VEP ANNOTATION OF *cis*-eQTLs OF CORE ADME GENES

Of the six *cis*-eQTLs identified for the core ADME genes, RegulomeDB identified only two *cis*-eQTLs, namely rs933765367, *cis*-eQTL for CYP2C19 and rs1000598999, *cis*-eQTL for SULT1A1, with scores greater than 0.5 (Table 3.29: RegulomeDB annotation for *cis-eQTLs* of core ADME genes). The score ranges from 0 to 1. A score of zero means that it not a regulatory variant and one means that it is most likely a regulatory variant. rs933765367, *cis*-eQTL for CYP2C19 had a score of 0.56761 and a rank of 3a. rs1000598999, *cis*-eQTL for SULT1A1 had a score of 0.77072 and a rank of 6. A rank of 3a refers to 'Transcription factor binding + any motif + DNase peak' and a rank of 6 refers to 'Motif hit' (https://regulomedb.org/regulome-help/). SNPs that show the strongest evidence of being regulatory in nature are given a rank of 1 and SNPs that show the least evidence of being functional are given a rank of 7. Therefore, rs933765367, *cis*-eQTL for CYP2C19 has stronger evidence of being a regulatory variant than rs1000598999, the *cis*-eQTL for SULT1A1. The *cis*-eQTL for CYP2C19, rs933765367, was also found to be associated with the chromatin state 'Enhancer' and 'Active Transcription Start Site'.

Of the six *cis*-eQTLs identified for the core ADME genes, VEP identified three as intronic variants (*cis*-eQTLs for *CYP2C19*, *UGT2B17* and *SULT1A1*), one as an upstream synonymous variant (*cis*-eQTL for *CYP3A5*), one as an intergenic variant (*cis*-eQTL for *DPYD*) and one as a downstream variant (*cis*-eQTL for *UGT1A1*) (Table 3.30).

4.16 PREVIOUSLY IDENTIFIED PHARMACOGENETIC VARIANTS OF ADME GENES OF RELEVANCE TO THIS STUDY

Pharmacogenetic variants such as *CYP2C19*17*, which is associated with increased *CYP2C19* expression resulting in an isozyme with increased enzyme activity, and *UGT1A1*28*, which is associated with decreased *UGT1A1* expression resulting in an isozyme with decreased enzyme activity, that have previously been identified to alter the expression of ADME genes, were not identified as eQTLs in this study. Correlating eQTLs with gene expression, relies on the source and integrity of the expression data. In the context of ADME genes, level of expression depends on the inducibility of expression in the presence of exposure to a drug

(Raucy *et al.*, 2002; Sumida *et al.*, 2013). In the absence of exposure, expression levels are low, but in the presence of and inducer of expression, such as exposure to a drug, expression levels increase dramatically. Furthermore, these, ADME genes can also be expressed in various tissue types for example, *UGT1A1* can be expressed in the liver, gastrointestinal tract, gall bladder, kidneys, or blood (https://www.proteinatlas.org/ENSG00000241635-UGT1A1/tissue) and *CYP2C19* can be expressed in the liver or the gastrointestinal tract (https://www.proteinatlas.org/ENSG00000165841-CYP2C19/tissue). Since expression levels in other tissues can differ to those in the liver, which was the organ focused upon in this study, data from non-liver focused studies cannot be directly correlated. This also highlights the limitations of such studies and the need to better characterise tissue sources.

4.17 LIMITATIONS OF THE STUDY

There are limited data in the public domain on individuals with both liver transcriptome and whole genome sequence data for the detection of eQTLs. At present we only know of one dataset available for individuals of African ancestry, and that dataset is of individuals of African-American ancestry (Zhong *et al.*, 2020). This study is therefore limited by the available data. As new data becomes available, the same pipeline can be used to refine the conclusions.

A major limitation of this study is that the pipeline was only run on one genotype-gene expression dataset pair of 149 samples (GSE39036-GSE32505) (Schröder *et al.*, 2013). Initially the intention was to use at least three datasets. The other two genotype-gene expression dataset pairs intended to be used for this study was a part of the study by Innocenti et al., 2011 (Innocenti *et al.*, 2011). The bigger genotype-gene expression dataset pair was GSE26105-GSE25935 comprising 224 samples. QC of genotype dataset (GSE26105) was successfully done. However, QC of the gene expression dataset (GSE25935) could not be done because almost all the samples included in this dataset had missing gene expression values. The smaller genotype-gene expression dataset pair consisted of 60 samples; the gene expression dataset of this pair is GSE28893. As the publication stated that both the genotype and gene expression datasets were available in GEO, QC of the gene expression dataset was first done. However, after performing gene expression QC, the corresponding genotype

dataset could not be found in GEO. When the author was contacted via email, he indicated that at the time of the study they did not have permission to release the genotype dataset and he has no information on what happened to the data after the study was published. Thus, instead of running the pipeline through a total of 433 samples, it was only run using one genotype-gene expression dataset pair comprising 149 samples.

Furthermore, no eQTLs were found from the genotype-gene expression dataset pair prior to imputation of the genotype dataset. When doing imputation, it is important to have a reference dataset that is from the same populations that the datasets are derived from to have accurate imputation, especially for rare variants.

The eQTLs identified for the core and extended ADME genes were not found in the GTEx database. They also could not be verified by the GTEx eQTL calculator. Each time an input was submitted, GTEx returned an unspecified error, necessitating further investigation. One explanation could be that the variant was not present in the GTEx database at the time the comparison was done or that the appropriate tissue was not adequately represented in GTEx. Although liver eQTLs in GTEx have been generated using a bigger dataset (226 samples of which genotype data is available for 208 samples) compared to the 136 liver samples from *Schröder et al.* 2013 that passed QC in this study, eQTLs identified by the pipeline were found to not be replicated in GTEx despite the samples being of similar European ancestry. This could be because of different QC methods employed by this study and GTEx. Differences in eQTLs identified for a gene could also arise because of differences in sample source. All the GTEx samples were obtained from cadavers whereas samples from *Schröder et al.* 2013 were obtained from people undergoing liver surgery. The effect of post-mortem time on samples obtained from cadavers might influence the expression of genes in the liver. The eQTLs were also not found to be present in QTLbase. Therefore, it is of utmost importance that the novel eQTLs identified in this study must be clinically and functionally validated.

Challenges faced in genotyping ADME genes are also a limiting factor in the identification of eQTLs. If a gene is not genotyped with high accuracy, it is impossible to predict its associated eQTLs. A well-known example of this is the *CYP2D6* gene (Cohn *et al.*, 2017). Although *CYP2D6* is relatively small in size, its polymorphic nature and loci surrounding the gene adds to the complexity of it being genotyped with high accuracy (Fang *et al.*, 2014).

Moreover, the liver is not a homogenous organ, it contains many different cell types including hepatocytes, blood vessel epithelium and blood. It was not stated in the original paper whether the 149 samples from the study were from liver sections, or a sub-set of specific types of cells.

4.18 FUTURE STUDIES

Since all the findings of this study are based on one genotype-gene expression dataset pair (due to unavailability of other datasets in the public domain), the findings of this analysis must be validated by additional studies. Furthermore, these results are based on bioinformatic analyses only and should be followed up by functional validation and assessments. Functional studies evaluating how eQTLs affect the dose at which drugs are prescribed must be done to identify optimal drug dosage levels for patients with and without *cis*-eQTLs.

The identification of regulatory variants that affect the expression of ADME genes in different tissues is an important field of research with implications for drug use and safety. This research is necessary not only in Europeans, but especially in African populations that have much more genetic diversity. African populations also have high genetic substructure (Sengupta *et al.*, 2020) and it would be important to investigate multiple ethnolinguistic groups across different geographic regions to understand mutation characteristics (occurrence and frequency) to be able to predict the impact of different genetic variants on the use of specific genes.

Though novel eQTLs in hepatocytes have been discovered by Zhong *et al.* in African-Americans, an admixed population with African ancestry, no studies have yet been done in livers on African populations resident in Africa (Zhong *et al.*, 2020). Since there is no data on liver eQTLs from African populations resident in Africa, there is a need to encourage and support the generation of complex datasets (whole genome genotyping or whole genome sequencing together with transcriptome data from different tissues) to develop an African liver eQTL database or an African eQTL database. As no such database exists at present, a database with a focus on populations resident in Africa would prove extremely useful, not only for detecting ADME gene related eQTLs, but also to inform genome-wide associations with complex traits and gene function more broadly to gain a deeper understanding of the molecular basis of diseases.

4.19 CONCLUSION

eQTLs and their direct association with tissue-specific gene expression could provide potential new links between the genotype and phenotype that could be clinically relevant. This information will enhance personalised medicine by providing vital information that can lead to the identification of new disease biomarkers, new therapies, and diagnostic procedures.

This study described the development of a pipeline to identify *cis*- and *trans*-eQTLs from a dataset with genotype and transcriptome data from the same individuals. The effect of the identified *cis*-eQTLs on drugs that are commonly prescribed in South Africa were explored and documented. If these eQTLs are functionally validated, their effects on drug metabolism and transport could be predicted and the information could be used to provide individual information that could, in turn, be used to reduce ADRs and to optimise treatment (e.g. dose). This could potentially significantly reduce the cost of treatment by reducing the risk for developing ADR. eQTLs identified in this study have been hypothesized to predict clinical outcomes of drug treatment like the occurrence of ADRs and the need to modify drug doses for better treatment outcomes, or to prescribe alternate therapies. If validated, treatment response predictions/algorithms could be enhanced by adding these eQTLs in combination with existing knowledge of genetic variations and other clinical factors. Developing

pharmacogenetic tests that include eQTLs in addition to existing well-known genetic variants can be a novel strategy to identify individuals who may be able to tolerate medication better as well as those who are more likely to achieve a therapeutic response or require alternative treatment. Thus, the focus of this study was on the potential role of eQTLs in pharmacogenomics.

Mapping of eQTLs in liver tissue has immense power and potential for understanding inter-individual variability in response to drug treatments. The eQTLs of ADME genes identified in this study can potentially explain some of the variance in drug dose requirements that contribute to the currently unexplainable portion of variability in drug response. ADME gene eQTLs in the liver can explain the associations of SNPs in these ADME genes with optimal drug dosage or with toxicity to an extent. This provides a sound rationale for using ADME gene-eQTL liver data for candidate SNP selection for pharmacogenomic study.

The *cis*-eQTLs identified in this study were found to be present at very different frequencies in African populations compared to other populations. This highlights the importance of doing more eQTL studies with samples of African origin to understand differences pertaining to drug pharmacogenetics. This is of importance especially for drugs that are administered as antiretroviral therapies that are used extensively in sub-Saharan African countries due to the very high disease burden of HIV and AIDS. Understanding what effect *cis*-eQTLs have on the metabolism of these drugs is of utmost importance to attain optimal drug efficacy levels.

The bioinformatic pipeline developed in this study therefore will be a useful resource for researchers to identify eQTLs associated with any kind of tissue. The only pre-requisite is availability of genotype data and corresponding normalised gene expression data from the same set of individuals.

REFERENCES

Achenbach, C. J. *et al.* (2011) 'Atazanavir/ritonavir-based combination antiretroviral therapy for treatment of HIV-1 infection in adults', *Future virology*, 6(2), pp. 157–177. doi: 10.2217/fvl.10.89.

Aguet, F. *et al.* (2017) 'Genetic effects on gene expression across human tissues', *Nature*, 550(7675), pp. 204–213. doi: 10.1038/nature24277.

Aldrich, S. L. *et al.* (2019) 'Influence of CYP2C19 Metabolizer Status on Escitalopram/Citalopram Tolerability and Response in Youth With Anxiety and Depressive Disorders', *Frontiers in pharmacology*, 10, p. 99. doi: 10.3389/fphar.2019.00099.

Alharbi, K. K. *et al.* (2013) 'Insertion/Deletion polymorphisms do play any role in G6PD deficiency individuals in the Kingdom of the Saudi Arabia', *Bioinformation*. 2013/01/09, 9(1), pp. 49–53. doi: 10.6026/97320630009049.

Alharbi, K. K. (2015) 'Genetic polymorphisms in paraoxonase 1 and G protein-coupled receptor 77, and the risk of glucose-6-phosphate dehydrogenase deficiency in a Saudi population', *Saudi medical journal*, 36(5), pp. 544–548. doi: 10.15537/smj.2015.5.11860.

Ali, Z., Kim, R. and Ysla, F. (2009) 'CYP2C9 polymorphisms: considerations in NSAID therapy', *Current opinion in drug discovery & development*, 12, pp. 108–114.

Altshuler, D., Daly, M. J. and Lander, E. S. (2008) 'Genetic Mapping in Human Disease', *Science*, 322(5903), pp. 881 LP – 888. doi: 10.1126/science.1156409.

Amstutz, U. *et al.* (2009) 'Dihydropyrimidine dehydrogenase gene variation and severe 5-fluorouracil toxicity: a haplotype assessment', *Pharmacogenomics*, 10(6), pp. 931–944. doi: 10.2217/pgs.09.28.

Amstutz, U. *et al.* (2018) 'Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing: 2017 Update', *Clinical pharmacology and therapeutics*. 2017/11/20, 103(2), pp. 210–216. doi: 10.1002/cpt.911.

Anderle, A. *et al.* (2018) 'Point-of-Care Testing for G6PD Deficiency: Opportunities for Screening', *International journal of neonatal screening*. 2018/11/19, 4(4), p. 34. doi:

10.3390/ijns4040034.

Anderson, C. A. *et al.* (2010) 'Data quality control in genetic case-control association studies', *Nature protocols*. 2010/08/26, 5(9), pp. 1564–1573. doi: 10.1038/nprot.2010.116.

Anderson, P. L. *et al.* (2009) 'Atazanavir pharmacokinetics in genetically determined CYP3A5 expressors versus non-expressors', *The Journal of antimicrobial chemotherapy*. 2009/08/26, 64(5), pp. 1071–1079. doi: 10.1093/jac/dkp317.

Arbitrio, M. *et al.* (2016) 'DMET™ (Drug Metabolism Enzymes and Transporters): A Pharmacogenomic platform for precision medicine', *Oncotarget*, 7(33), pp. 54028–54050. doi: 10.18632/oncotarget.9927.

Baccarani, M. *et al.* (2009) 'Chronic myeloid leukemia: an update of concepts and management recommendations of European LeukemiaNet', *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2009/11/02, 27(35), pp. 6041–6051. doi: 10.1200/JCO.2009.25.0779.

Bader, L. A. and Elewa, H. (2016) 'The Impact of Genetic and Non-Genetic Factors on Warfarin Dose Prediction in MENA Region: A Systematic Review', *PloS one*, 11(12), pp. e0168732–e0168732. doi: 10.1371/journal.pone.0168732.

Bains, R. K. (2013) 'African variation at Cytochrome P450 genes: Evolutionary aspects and the implications for the treatment of infectious diseases', *Evolution, medicine, and public health*. 2013/05/15, 2013(1), pp. 118–134. doi: 10.1093/emph/eot010.

Bairam, A. F. *et al.* (2017) 'On the Molecular Basis Underlying the Metabolism of Tapentadol Through Sulfation', *European journal of drug metabolism and pharmacokinetics*, 42(5), p. 793—800. doi: 10.1007/s13318-016-0392-8.

Baker, D. E. (2003) 'Pharmacogenomics of azathioprine and 6-mercaptopurine in gastroenterologic therapy', *Reviews in Gastroenterological Disorders*, 3(3), pp. 150–157.

Balram, C. *et al.* (2003) 'CYP3A5*3 and *6 single nucleotide polymorphisms in three distinct Asian populations', *European Journal of Clinical Pharmacology*, 59(2), pp. 123–126. doi: 10.1007/s00228-003-0594-2.

Barbarino, J. M. *et al.* (2014) 'PharmGKB summary: very important pharmacogene

information for UGT1A1', *Pharmacogenetics and genomics*, 24(3), pp. 177–183. doi: 10.1097/FPC.0000000000000024.

Barratt, D. T. and Somogyi, A. A. (2017) 'Role of pharmacogenetics in personalised imatinib dosing', *Translational Cancer Research*, 6(Suppl 10), pp. S1541–S1557. doi: 10.21037/tcr.2017.09.08.

Barrett, J. C. *et al.* (2008) 'Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease', *Nature genetics*. 2008/06/29, 40(8), pp. 955–962. doi: 10.1038/ng.175.

Bathum, L. *et al.* (1999) 'Phenotypes and genotypes for CYP2D6 and CYP2C19 in a black Tanzanian population', *British journal of clinical pharmacology*, 48(3), pp. 395–401. doi: 10.1046/j.1365-2125.1999.00019.x.

Belle, D. J. and Singh, H. (2008) 'Genetic factors in drug metabolism', *American Family Physician*, 77(11), pp. 1553–1560.

Bennett, J. W. *et al.* (2013) 'Primaquine Failure and Cytochrome P-450 2D6 in Plasmodium vivax Malaria', *New England Journal of Medicine*, 369(14), pp. 1381–1382. doi: 10.1056/NEJMc1301936.

Bertilsson, L. (1995) 'Geographical/Interracial Differences in Polymorphic Drug Oxidation', *Clinical Pharmacokinetics*, 29(3), pp. 192–209. doi: 10.2165/00003088-199529030-00005.

Beutler, E. (1991) 'Glucose-6-Phosphate Dehydrogenase Deficiency', *New England Journal of Medicine*, 324(3), pp. 169–174. doi: 10.1056/NEJM199101173240306.

Beutler, E. (1994) 'G6PD deficiency', *Blood*, 84(11), pp. 3613–3636. doi: 10.1182/blood.V84.11.3613.bloodjournal84113613.

Beutler, E., Gelbart, T. and Demina, A. (1998) 'Racial variability in the UDP-glucuronosyltransferase 1 (UGT1A1) promoter: a balanced polymorphism for regulation of bilirubin metabolism?', *Proceedings of the National Academy of Sciences of the United States of America*, 95(14), pp. 8170–8174. doi: 10.1073/pnas.95.14.8170.

BEUTLER, E., YEH, M. and FAIRBANKS, V. F. (1962) 'The normal human female as a mosaic of X-chromosome activity: studies using the gene for C-6-PD-deficiency as a marker',

Proceedings of the National Academy of Sciences of the United States of America, 48(1), pp. 9–16. doi: 10.1073/pnas.48.1.9.

Birdwell, K. A. *et al.* (2015) 'Clinical Pharmacogenetics Implementation Consortium (CPIC) Guidelines for CYP3A5 Genotype and Tacrolimus Dosing', *Clinical pharmacology and therapeutics*. 2015/06/03, 98(1), pp. 19–24. doi: 10.1002/cpt.113.

Biswas, M. *et al.* (2018) 'Association between genetic polymorphisms of CYP2C9 and VKORC1 and safety and efficacy of warfarin: Results of a 5 years audit', *Indian heart journal*. 2018/03/01, 70 Suppl 3(Suppl 3), pp. S13–S19. doi: 10.1016/j.ihj.2018.02.005.

Blume, H. *et al.* (2006) 'Pharmacokinetic Drug Interaction Profiles of Proton Pump Inhibitors', *Drug Safety*, 29(9), pp. 769–784. doi: 10.2165/00002018-200629090-00002.

Bodin, L. *et al.* (2005) 'Cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase (VKORC1) genotypes as determinants of acenocoumarol sensitivity', *Blood*, 106(1), pp. 135–140. doi: 10.1182/blood-2005-01-0341.

Bogari, N. M. (2016) 'Next generation sequencing (NGS) in glucose-6-phosphate dehydrogenase (G6PD) deficiency studies', *Bioinformation*, 12(2), pp. 41–43. doi: 10.6026/97320630012041.

Boige, V. *et al.* (2016) 'DPYD Genotyping to Predict Adverse Events Following Treatment With Fluorouracil-Based Adjuvant Chemotherapy in Patients With Stage III Colon Cancer: A Secondary Analysis of the PETACC-8 Randomized Clinical Trial', *JAMA Oncology*, 2(5), pp. 655–662. doi: 10.1001/jamaoncol.2015.5392.

Bonifaz-Peña, V. *et al.* (2014) 'Exploring the distribution of genetic markers of pharmacogenomics relevance in Brazilian and Mexican populations', *PLoS ONE*, 9(11), p. 122161. doi: 10.1371/journal.pone.0112640.

Van Booven, D. *et al.* (2010) 'Cytochrome P450 2C9-CYP2C9', *Pharmacogenetics and genomics*, 20(4), pp. 277–281. doi: 10.1097/FPC.0b013e3283349e84.

Börnigen, D. and Zeller, T. (2018) 'Transcriptomics in Cardiovascular Medicine', in Vasan, R. S. and Sawyer, D. B. B. T.-E. of C. R. and M. (eds). Oxford: Elsevier, pp. 558–571. doi: https://doi.org/10.1016/B978-0-12-809657-4.99572-3.

Borobia, A. M. *et al.* (2009) 'Trough Tacrolimus Concentrations in the First Week After Kidney Transplantation Are Related to Acute Rejection', *Therapeutic Drug Monitoring*, 31(4). Available at: https://journals.lww.com/drug-monitoring/Fulltext/2009/08000/Trough_Tacrolimus_Concentrations_in_the_First_Week.3.aspx.

Boyle, A. P. *et al.* (2012) 'Annotation of functional variation in personal genomes using RegulomeDB', *Genome research*, 22(9), pp. 1790–1797. doi: 10.1101/gr.137323.112.

Brand, M., Gaylard, P. and Ramos, J. (2018) 'Colorectal cancer in South Africa: An assessment of disease presentation, treatment pathways and 5-year survival', *South African Medical Journal*, 108(2), p. 118. doi: 10.7196/SAMJ.2018.v108i2.12338.

Brem, R. B. and Kruglyak, L. (2005) 'The landscape of genetic complexity across 5,700 gene expression traits in yeast', *Proceedings of the National Academy of Sciences of the United States of America*, 102(5), pp. 1572 LP – 1577. doi: 10.1073/pnas.0408709102.

Bright, A. T. *et al.* (2013) 'Genetic analysis of primaquine tolerance in a patient with relapsing vivax malaria', *Emerging infectious diseases*, 19(5), pp. 802–805. doi: 10.3201/eid1905.121852.

Brittain, H. K., Scott, R. and Thomas, E. (2017) 'The rise of the genome and personalised medicine', *Clinical medicine (London, England)*, 17(6), pp. 545–551. doi: 10.7861/clinmedicine.17-6-545.

Brown, K. C. *et al.* (2011) 'Single and multiple dose pharmacokinetics of maraviroc in saliva, semen, and rectal tissue of healthy HIV-negative men', *The Journal of infectious diseases*, 203(10), pp. 1484–1490. doi: 10.1093/infdis/jir059.

Buniello, A. *et al.* (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Research*, 47(D1), pp. D1005–D1012. doi: 10.1093/nar/gky1120.

Bura, A. *et al.* (2006) 'Cytochrome P450 2C19 loss-of-function polymorphism is a major determinant of clopidogrel responsiveness in healthy subjects', *Blood*, 108(7), pp. 2244–2247. doi: 10.1182/blood-2006-04-013052.Supported.

Campbell, M. C. and Tishkoff, S. A. (2008) 'African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping', *Annual review of genomics and human genetics*, 9, pp. 403–433. doi: 10.1146/annurev.genom.9.081307.164258.

Campbell, M. C. and Tishkoff, S. A. (2010) 'The Evolution of Human Genetic and Phenotypic Variation in Africa', *Current Biology*, 20(4), pp. R166–R173. doi: https://doi.org/10.1016/j.cub.2009.11.050.

Cao, X. (2020) 'Joint variable selection and network modeling for detecting eQTLs', pp. 1–15. doi: 10.1515/sagmb-2019-0032.

Caraco, Y., Blotnick, S. and Muszkat, M. (2008) 'CYP2C9 Genotype-guided Warfarin Prescribing Enhances the Efficacy and Safety of Anticoagulation: A Prospective Randomized Controlled Study', *Clinical Pharmacology & Therapeutics*, 83(3), pp. 460–470. doi: 10.1038/sj.clpt.6100316.

Carlini, E. J. *et al.* (2001) 'Sulfation pharmacogenetics:SULT1A1 and SULT1A2 allele frequencies in Caucasian, Chinese and African-American subjects', *Pharmacogenetics and Genomics*, 11(1). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2001/02000/Sulfation_pharmacogenetics_SULT1A1_and_SULT1A2.7.aspx.

Carrasco-Ramiro, F., Peiró-Pastor, R. and Aguado, B. (2017) 'Human genomics projects and precision medicine', *Gene Therapy*, 24(9), pp. 551–561. doi: 10.1038/gt.2017.77.

CARSON, P. E. *et al.* (1956) 'Enzymatic Deficiency in Primaquine-Sensitive Erythrocytes', *Science*, 124(3220), pp. 484 LP – 485. doi: 10.1126/science.124.3220.484-a.

Castellino, S. *et al.* (2013) 'Metabolism, excretion, and mass balance of the HIV-1 integrase inhibitor dolutegravir in humans', *Antimicrobial agents and chemotherapy*. 2013/05/13, 57(8), pp. 3536–3546. doi: 10.1128/AAC.00292-13.

Caudle, K. E. *et al.* (2013) 'Clinical Pharmacogenetics Implementation Consortium Guidelines for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing', *Clinical Pharmacology & Therapeutics*, 94(6), pp. 640–645. doi: 10.1038/clpt.2013.172.

Ch Yiannakopoulou, E. (2013) 'Pharmacogenomics of phase II metabolizing enzymes and drug transporters: Clinical implications', *Pharmacogenomics Journal*, 13(2), pp. 105–109. doi: 10.1038/tpj.2012.42.

Chaichoompu, K. *et al.* (2020) 'A different view on fine-scale population structure in Western African populations', *Human Genetics*, 139(1), pp. 45–59. doi: 10.1007/s00439-019-02069-7.

Chang, K.-L., Weitzel, K. and Schmidt, S. (2015) 'Pharmacogenetics: Using Genetic Information to Guide Drug Therapy', *American family physician*, 92(7), pp. 588–594. Available at: https://pubmed.ncbi.nlm.nih.gov/26447442.

Chang, M. *et al.* (2014) 'Impact of Cytochrome P450 2C19 Polymorphisms on Citalopram/Escitalopram Exposure: A Systematic Review and Meta-Analysis', *Clinical Pharmacokinetics*, 53(9), pp. 801–811. doi: 10.1007/s40262-014-0162-1.

Chanhom, N. *et al.* (2020) 'GSTM1 and GSTT1 genetic polymorphisms and their association with antituberculosis drug-induced liver injury', *Biomedical reports*. 2020/02/10, 12(4), pp. 153–162. doi: 10.3892/br.2020.1275.

Chenevix-Trench, G. *et al.* (1995) 'Glutathione S-transferase M1 and T1 polymorphisms: susceptibility to colon cancer and age of onset', *Carcinogenesis*, 16(7), pp. 1655–1657. doi: 10.1093/carcin/16.7.1655.

Choudhury, A. *et al.* (2018) 'African genetic diversity provides novel insights into evolutionary history and local adaptations', *Human Molecular Genetics*, 27(R2), pp. R209–R218. doi: 10.1093/hmg/ddy161.

Choudhury, A. *et al.* (2020) 'High-depth African genomes inform human migration and health', *Nature*, 586(7831), pp. 741–748. doi: 10.1038/s41586-020-2859-7.

Chowbay, B., Zhou, S. and Lee, E. J. D. (2005) 'An Interethnic Comparison of Polymorphisms of the Genes Encoding Drug-Metabolizing Enzymes and Drug Transporters: Experience in Singapore', *Drug Metabolism Reviews*, 37(2), pp. 327–378. doi: 10.1081/DMR-28805.

Ciccacci, C. *et al.* (2009) 'Nevirapine-induced hepatotoxicity and pharmacogenetics: a retrospective study in a population from Mozambique', *Pharmacogenomics*, 11(1), pp. 23–

31. doi: 10.2217/pgs.09.142.

Cleveland Clinic Laboratories (2013) 'Genetic Test of Cytochrome P450 2C19 (CYP2C19) for Drug Metabolism', 19(Table 2). Available at: http://portals.clevelandclinic.org/portals/66/pdf/technical-briefs/genetic-test-of-cytochrome-p450-2c19-cyp2c19-for-drug-metabolism.pdf.

Cohn, I. *et al.* (2017) 'Genome sequencing as a platform for pharmacogenetic genotyping: a pediatric cohort study', *npj Genomic Medicine*, 2(1), p. 19. doi: 10.1038/s41525-017-0021-8.

Consortium, Gte. (2013) 'The Genotype-Tissue Expression (GTEx) project', *Nature genetics*, 45(6), pp. 580–585. doi: 10.1038/ng.2653.

Consortium, T. Gte. (2015) 'The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans', *Science*, 348(6235), pp. 648 LP – 660. doi: 10.1126/science.1262110.

Consortium, W. T. C. C. (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447(7145), pp. 661–678. doi: 10.1038/nature05911.

Consortium, W. T. C. C. *et al.* (2010) 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls', *Nature*, 464(7289), pp. 713–720. doi: 10.1038/nature08979.

Cooper, G. M., Johnson, J. A., *et al.* (2008) 'A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose', *Blood*. 2008/06/05, 112(4), pp. 1022–1027. doi: 10.1182/blood-2008-01-134247.

Cooper, G. M., Zerr, T., *et al.* (2008) 'Systematic assessment of copy-number variant detection via genome-wide single nucleotide polymorphism genotyping', *Nature genetics*, 40(10), pp. 1199–1203. doi: 10.1038/ng.236.Systematic.

Cullell, N. *et al.* (2018) 'Pharmacogenetic studies with oral anticoagulants. Genome-wide association studies in vitamin K antagonist and direct oral anticoagulants', *Oncotarget; Vol 9, No 49*. Available at: https://www.oncotarget.com/article/25579/text/.

Cuzin, L. *et al.* (2019) 'Integrase strand transfer inhibitors and neuropsychiatric adverse

events in a large prospective cohort', *Journal of Antimicrobial Chemotherapy*, 74(3), pp. 754–760. doi: 10.1093/jac/dky497.

Czaja, A. S., Valuck, R. J. and Anderson, H. D. (2013) 'Comparative safety of selective serotonin reuptake inhibitors among pediatric users with respect to adverse cardiac events', *Pharmacoepidemiology and Drug Safety*, 22(6), pp. 607–614. doi: 10.1002/pds.3420.

Dai, X. *et al.* (2013) 'A Genome-Wide Association Study for Serum Bilirubin Levels and Gene-Environment Interaction in a Chinese Population', *Genetic Epidemiology*, 37(3), pp. 293–300. doi: 10.1002/gepi.21711.

Dai, Y. *et al.* (2004) 'In vitro metabolism of cyclosporine A by human kidney CYP3A5', *Biochemical Pharmacology*, 68(9), pp. 1889–1902. doi: https://doi.org/10.1016/j.bcp.2004.07.012.

Daly, A. K. (2006) 'Significance of the Minor Cytochrome P450 3A Isoforms', *Clinical Pharmacokinetics*, 45(1), pp. 13–31. doi: 10.2165/00003088-200645010-00002.

Daly, A. K. and King, B. P. (2003) 'Pharmacogenetics of oral anticoagulants', *Pharmacogenetics and Genomics*, 13(5). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2003/05000/Pharmacogenetics_of_oral_anticoagulants.2.aspx.

Dandara, C. *et al.* (2001) 'Genetic polymorphism of CYP2D6 and CYP2C19 in East- and Southern African populations including psychiatric patients', *European Journal of Clinical Pharmacology*, 57(1), pp. 11–17. doi: 10.1007/s002280100282.

Dandara, C., Ballo, R. and Iqbal Parker, M. (2005) 'CYP3A5 genotypes and risk of oesophageal cancer in two South African populations', *Cancer Letters*, 225(2), pp. 275–282. doi: https://doi.org/10.1016/j.canlet.2004.11.004.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics*, 27(15), pp. 2156–2158. doi: 10.1093/bioinformatics/btr330.

Dang, M.-T. N., Hambleton, J. and Kayser, S. R. (2005) 'The Influence of Ethnicity on Warfarin Dosage Requirement', *Annals of Pharmacotherapy*, 39(6), pp. 1008–1012. doi: 10.1345/aph.1E566.

Danoff, T. M. *et al.* (2004) 'A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia', *The Pharmacogenomics Journal*, 4(1), pp. 49–53. doi: 10.1038/sj.tpj.6500221.

Davidson, D. G. and Eastham, W. N. (1966) 'Acute liver necrosis following overdose of paracetamol', *British medical journal*, 2(5512), pp. 497–499. doi: 10.1136/bmj.2.5512.497.

Dayem Ullah, A. Z., Lemoine, N. R. and Chelala, C. (2012) 'SNPnexus: A web server for functional annotation of novel and publicly known genetic variants (2012 update)', *Nucleic Acids Research*, 40(W1), pp. 65–70. doi: 10.1093/nar/gks364.

Dean, L. (2016) 'Capecitabine Therapy and DPYD Genotype Drug Class : Fluoropyrimidines Drug : Capecitabine', *Medical Genetics Summaries - NCBI Bookshelf*, (Md), pp. 1–7.

Dean, L. (2018) 'Irinotecan Therapy and UGT1A1 Genotype Drug : Irinotecan Irinotecan is used to treat colorectal cancer , which is the third most common cancer worldwide (6). It is often', (MD), pp. 1–12.

Deelen, P. *et al.* (2015) 'Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels', *Genome Medicine*, 7(1), p. 30. doi: 10.1186/s13073-015-0152-4.

Deenen, M. J. *et al.* (2011) 'Relationship between Single Nucleotide Polymorphisms and Haplotypes in DPYD and Toxicity and Efficacy of Capecitabine in Advanced Colorectal Cancer', *Clinical Cancer Research*, 17(10), pp. 3455 LP – 3468. doi: 10.1158/1078-0432.CCR-10-2209.

Deenen, M. J. *et al.* (2015) 'Upfront Genotyping of DPYD*2A to Individualize Fluoropyrimidine Therapy: A Safety and Cost Analysis', *Journal of Clinical Oncology*, 34(3), pp. 227–234. doi: 10.1200/JCO.2015.63.1325.

Dias, M. M., McKinnon, R. A. and Sorich, M. J. (2012) 'Impact of the UGT1A1*28 allele on response to irinotecan: a systematic review and meta-analysis', *Pharmacogenomics*, 13(8), pp. 889–899. doi: 10.2217/pgs.12.68.

Ding, C. *et al.* (2016) 'A Cell-type-resolved Liver Proteome', *Molecular & cellular proteomics : MCP*. 2016/08/25, 15(10), pp. 3190–3202. doi: 10.1074/mcp.M116.060145.

DiPiero, J., Teng, K. and Hicks, J. K. (2015) 'Should thiopurine methyltransferase (TPMT) activity be determined before prescribing azathioprine, mercaptopurine, or thioguanine?', *Cleveland Clinic Journal of Medicine*, 82(7), pp. 409 LP – 413. doi: 10.3949/ccjm.82a.14106.

Dorr, P. *et al.* (2005) 'Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity', *Antimicrobial agents and chemotherapy*, 49(11), pp. 4721–4732. doi: 10.1128/AAC.49.11.4721-4732.2005.

Dreisbach, A. W. *et al.* (2005) 'The Prevalence of CYP2C8, 2C9, 2J2, and Soluble Epoxide Hydrolase Polymorphisms in African Americans With Hypertension*', *American Journal of Hypertension*, 18(10), pp. 1276–1281. doi: 10.1016/j.amjhyper.2005.04.019.

Dulucq, S. and Krajinovic, M. (2010) 'The pharmacogenetics of imanitib', *Genome medicine*, 2(11), p. 85. doi: 10.1186/gm206.

Eap, C. B. *et al.* (2004) 'Nonresponse to Clozapine and Ultrarapid CYP1A2 Activity: Clinical Data and Analysis of CYP1A2 Gene', *Journal of Clinical Psychopharmacology*, 24(2). Available at: https://journals.lww.com/psychopharmacology/Fulltext/2004/04000/Nonresponse_to_Clozapine_and_Ultrarapid_CYP1A2.16.aspx.

Eckford, P. D. W. *et al.* (2012) 'Cystic fibrosis transmembrane conductance regulator (CFTR) potentiator VX-770 (ivacaftor) opens the defective channel gate of mutant CFTR in a phosphorylation-dependent but ATP-independent manner', *The Journal of biological chemistry*. 2012/08/31, 287(44), pp. 36639–36649. doi: 10.1074/jbc.M112.393637.

Egbelakin, A. *et al.* (2011) 'Increased risk of vincristine neurotoxicity associated with low CYP3A5 expression genotype in children with acute lymphoblastic leukemia', *Pediatric Blood & Cancer*, 56(3), pp. 361–367. doi: 10.1002/pbc.22845.

Ehmann, F. *et al.* (2015) 'Pharmacogenomic information in drug labels: European Medicines Agency perspective', *The Pharmacogenomics Journal*, 15(3), pp. 201–210. doi: 10.1038/tpj.2014.86.

Ekhart, C., Rodenhuis, S., *et al.* (2009) 'An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug

treatment', *Cancer Treatment Reviews*, 35(1), pp. 18–31. doi: 10.1016/j.ctrv.2008.07.003.

Ekhart, C., Doodeman, V. D., *et al.* (2009) 'Polymorphisms of drug-metabolizing enzymes (GST, CYP2B6 and CYP3A) affect the pharmacokinetics of thiotepa and tepa', *British Journal of Clinical Pharmacology*, 67(1), pp. 50–60. doi: 10.1111/j.1365-2125.2008.03321.x.

Elzi, L. *et al.* (2017) 'Adverse events of raltegravir and dolutegravir', *AIDS (London, England)*, 31(13), pp. 1853–1858. doi: 10.1097/QAD.0000000000001590.

Emslie, G. J. *et al.* (2009) 'Escitalopram in the Treatment of Adolescent Depression: A Randomized Placebo-Controlled Multisite Trial', *Journal of the American Academy of Child & Adolescent Psychiatry*, 48(7), pp. 721–729. doi: 10.1097/CHI.0b013e3181a2b304.

EVANS, D. A., MANLEY, K. A. and McKUSICK, V. A. (1960) 'Genetic control of isoniazid metabolism in man', *British medical journal*, 2(5197), pp. 485–491. doi: 10.1136/bmj.2.5197.485.

Evans, W. E. and Relling, M. V (1999) 'Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics', *Science*, 286(5439), pp. 487 LP – 491. doi: 10.1126/science.286.5439.487.

Falany, C. N., Ström, P. and Swedmark, S. (2005) 'Sulphation of o-desmethylnaproxen and related compounds by human cytosolic sulfotransferases', *British journal of clinical pharmacology*, 60(6), p. 632—640. doi: 10.1111/j.1365-2125.2005.02506.x.

Fang, H. *et al.* (2014) 'Establishment of CYP2D6 reference samples by multiple validated genotyping platforms', *The Pharmacogenomics Journal*, 14(6), pp. 564–572. doi: 10.1038/tpj.2014.27.

Ferguson, C. S. and Tyndale, R. F. (2011) 'Cytochrome P450 enzymes in the brain: emerging evidence of biological significance', *Trends in pharmacological sciences*. 2011/10/03, 32(12), pp. 708–714. doi: 10.1016/j.tips.2011.08.005.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) 'Structural variation in the human genome', 7(February), pp. 85–97. doi: 10.1038/nrg1767.

Franke, L. and Jansen, R. C. (2009) *eQTL analysis in humans.*, *Methods in molecular biology (Clifton, N.J.)*. Humana Press. doi: 10.1007/978-1-60761-247-6_17.

Froehlich, T. K. *et al.* (2015) 'Clinical importance of risk variants in the dihydropyrimidine dehydrogenase gene for the prediction of early-onset fluoropyrimidine toxicity', *International Journal of Cancer*, 136(3), pp. 730–739. doi: 10.1002/ijc.29025.

Fröhlich, M. *et al.* (2004) 'Association of the CYP3A5 A6986G (CYP3A5*3) polymorphism with saquinavir pharmacokinetics', *British Journal of Clinical Pharmacology*, 58(4), pp. 443–444. doi: 10.1111/j.1365-2125.2004.02159.x.

Furuta, T. *et al.* (2004) 'Pharmacogenomics of proton pump inhibitors', *Pharmacogenomics*, 5(2), pp. 181–202. doi: 10.1517/phgs.5.2.181.27483.

FURUTA, T. *et al.* (2005) 'Influence of CYP2C19 Pharmacogenetic Polymorphism on Proton Pump Inhibitor-based Therapies', *Drug Metabolism and Pharmacokinetics*, 20(3), pp. 153–167. doi: 10.2133/dmpk.20.153.

Gamazon, E. R. and Perera, M. (2012) 'Genome-wide approaches in pharmacogenomics: heritability estimation and pharmacoethnicity as primary challenges', *Pharmacogenomics*, 13(10), pp. 1101–1104. doi: 10.2217/pgs.12.88.

Gammal, R. S. *et al.* (2016) 'Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for UGT1A1 and Atazanavir Prescribing', *Clinical pharmacology and therapeutics*. 2015/11/09, 99(4), pp. 363–369. doi: 10.1002/cpt.269.

Garcia-Donas, J. *et al.* (2011) 'Single nucleotide polymorphism associations with response and toxic effects in patients with advanced renal-cell carcinoma treated with first-line sunitinib: a multicentre, observational, prospective study', *The Lancet Oncology*, 12(12), pp. 1143–1150. doi: https://doi.org/10.1016/S1470-2045(11)70266-2.

García-Martín, E. *et al.* (2006) 'Interethnic and Intraethnic Variability of CYP2C8 and CYP2C9 Polymorphisms in Healthy Individuals', *Molecular Diagnosis & Therapy*, 10(1), pp. 29–40. doi: 10.1007/BF03256440.

Gardiner, S. J. and Begg, E. J. (2006) 'Pharmacogenetics, Drug-Metabolizing Enzymes, and Clinical Practice', *Pharmacological Reviews*, 58(3), pp. 521 LP – 590. doi: 10.1124/pr.58.3.6.

Garrod, A. (1902) 'THE INCIDENCE OF ALKAPTONURIA : A STUDY IN CHEMICAL INDIVIDUALITY.', *The Lancet*, 160(4137), pp. 1616–1620. doi:

https://doi.org/10.1016/S0140-6736(01)41972-6.

GEARRY, R. B. and BARCLAY, M. L. (2005) 'Azathioprine and 6-mercaptopurine pharmacogenetics and metabolite monitoring in inflammatory bowel disease', *Journal of Gastroenterology and Hepatology*, 20(8), pp. 1149–1157. doi: 10.1111/j.1440-1746.2005.03832.x.

Ghotbi, R. *et al.* (2007) 'Comparisons of CYP1A2 genetic polymorphisms, enzyme activity and the genotype-phenotype relationship in Swedes and Koreans', *European Journal of Clinical Pharmacology*, 63(6), pp. 537–546. doi: 10.1007/s00228-007-0288-2.

Gilad, Y., Rifkin, S. A. and Pritchard, J. K. (2008) 'Revealing the architecture of gene regulation: the promise of eQTL studies', *Trends in genetics : TIG*. 2008/07/01, 24(8), pp. 408–415. doi: 10.1016/j.tig.2008.06.001.

Giral, H., Landmesser, U. and Kratzer, A. (2018) 'Into the Wild: GWAS Exploration of Non-coding RNAs', *Frontiers in cardiovascular medicine*, 5, p. 181. doi: 10.3389/fcvm.2018.00181.

Glubb, D. M., Dholakia, N. and Innocenti, F. (2012) 'Liver expression quantitative trait loci: a foundation for pharmacogenomic research', *Frontiers in genetics*, 3, p. 153. doi: 10.3389/fgene.2012.00153.

'Glucose-6-phosphate dehydrogenase deficiency. WHO Working Group' (1989) *Bulletin of the World Health Organization*, 67(6), pp. 601–611. Available at: http://europepmc.org/abstract/MED/2633878.

Godoy, W. *et al.* (2002) 'CYP2A6/2A7 and CYP2E1 expression in human oesophageal mucosa: Regional and inter-individual variation in expression and relevance to nitrosamine metabolism', *Carcinogenesis*, 23(4), pp. 611–616. doi: 10.1093/carcin/23.4.611.

Gold, H. T. *et al.* (2009) 'Cost effectiveness of pharmacogenetic testing for uridine diphosphate glucuronosyltransferase 1A1 before irinotecan administration for metastatic colorectal cancer', *Cancer*, 115(17), pp. 3858–3867. doi: 10.1002/cncr.24428.

Göring, H. H. H. *et al.* (2007) 'Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes', *Nature Genetics*, 39(10), pp. 1208–1216. doi: 10.1038/ng2119.

Gotta, V. *et al.* (2014) 'Large-scale imatinib dose–concentration–effect study in CML patients under routine care conditions', *Leukemia Research*, 38(7), pp. 764–772. doi: https://doi.org/10.1016/j.leukres.2014.03.023.

Gudbjartsson, D. F. *et al.* (2008) 'Many sequence variants affecting diversity of adult human height', *Nature Genetics*, 40(5), pp. 609–615. doi: 10.1038/ng.122.

Gunes, A. and Dahl, M.-L. (2008) 'Variation in CYP1A2 activity and its clinical implications: influence of environmental factors and genetic polymorphisms', *Pharmacogenomics*, 9(5), pp. 625–637. doi: 10.2217/14622416.9.5.625.

Gurdasani, D. *et al.* (2015) 'The African Genome Variation Project shapes medical genetics in Africa', *Nature*, 517(7534), pp. 327–332. doi: 10.1038/nature13997.

Haas, D. W. *et al.* (2004) 'Pharmacogenetics of efavirenz and central nervous system side effects: an Adult AIDS Clinical Trials Group study', *AIDS*, 18(18). Available at: https://journals.lww.com/aidsonline/Fulltext/2004/12030/Pharmacogenetics_of_efavirenz_and_central_nervous.6.aspx.

Hagymási, K. *et al.* (2011) 'Update on the pharmacogenomics of proton pump inhibitors', *Pharmacogenomics*, 12(6), pp. 873–888. doi: 10.2217/pgs.11.4.

Hakonarson, H. *et al.* (2007) 'A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene', *Nature*, 448(7153), pp. 591–594. doi: 10.1038/nature06010.

Han, J.-Y. *et al.* (2006) 'Comprehensive Analysis of UGT1A Polymorphisms Predictive for Pharmacokinetics and Treatment Outcome in Patients With Non–Small-Cell Lung Cancer Treated With Irinotecan and Cisplatin', *Journal of Clinical Oncology*, 24(15), pp. 2237–2244. doi: 10.1200/JCO.2005.03.0239.

Henricks, L. M. *et al.* (2015) 'Translating DPYD genotype into DPD phenotype: using the DPYD gene activity score', *Pharmacogenomics*, 16(11), pp. 1275–1284. doi: 10.2217/pgs.15.70.

Henricks, L. M. *et al.* (2017) 'DPYD genotype-guided dose individualization to improve patient safety of fluoropyrimidine therapy: call for a drug label update', *Annals of Oncology*, 28(12), pp. 2915–2922. doi: https://doi.org/10.1093/annonc/mdx411.

Henry, S. *et al.* (2021) 'Characterization of Gene Expression Signatures for the Identification of Cellular Heterogeneity in the Developing Mammary Gland', *Journal of Mammary Gland Biology and Neoplasia*, 26(1), pp. 43–66. doi: 10.1007/s10911-021-09486-3.

Herbert, H. *et al.* (2005) 'Association Between the CYP3A5 Genotype and Blood Pressure', *Hypertension*, 45(2), pp. 294–298. doi: 10.1161/01.HYP.0000151361.31736.96.

Herbst, C. *et al.* (2020) 'Colorectal Cancer (CRC) treatment and associated costs in the public sector compared to the private sector in Johannesburg, South Africa', *BMC Health Services Research*, 20(1), p. 290. doi: 10.1186/s12913-020-05112-w.

Hernandez, W. *et al.* (2017) 'Integrated analysis of genetic variation and gene expression reveals novel variant for increased warfarin dose requirement in African Americans', *Journal of thrombosis and haemostasis : JTH*. 2017/03/25, 15(4), pp. 735–743. doi: 10.1111/jth.13639.

Hicks, J. K. *et al.* (2017) 'Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update', *Clinical pharmacology and therapeutics*. 2017/02/13, 102(1), pp. 37–44. doi: 10.1002/cpt.597.

Hildebrandt, M. *et al.* (2009) 'Very important pharmacogene summary: sulfotransferase 1A1', *Pharmacogenetics and Genomics*, 19(6). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2009/06000/Very_important_pharmacogene_summary_.2.aspx.

Hillman, M. A. *et al.* (2005) 'A Prospective, Randomized Pilot Trial of Model-Based Warfarin Dose Initiation using CYP2C9 Genotype and Clinical Data', *Clinical Medicine & Research* , 3(3), pp. 137–145. doi: 10.3121/cmr.3.3.137.

Hindorff, L. A. *et al.* (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proceedings of the National Academy of Sciences*, 106(23), pp. 9362 LP – 9367. doi: 10.1073/pnas.0903103106.

Hiratsuka, M. *et al.* (2002) 'Allele and genotype frequencies of CYP2B6 and CYP3A5 in the Japanese population', *European Journal of Clinical Pharmacology*, 58(6), pp. 417–421. doi: 10.1007/s00228-002-0499-5.

Hom, G. *et al.* (2008) 'Association of Systemic Lupus Erythematosus with C8orf13–BLK and ITGAM–ITGAX', *New England Journal of Medicine*, 358(9), pp. 900–909. doi: 10.1056/NEJMoa0707865.

Hovelson, D. H. *et al.* (2017) 'Characterization of ADME gene variation in 21 populations by exome sequencing', *Pharmacogenetics and Genomics*, 27(3), pp. 89–100. doi: 10.1097/FPC.0000000000000260.

Hu, D. G. *et al.* (2019) 'Deregulation of the Genes that Are Involved in Drug Absorption, Distribution, Metabolism, and Excretion in Hepatocellular Carcinoma', *The Journal of pharmacology and experimental therapeutics*, 368(3), pp. 363–381. doi: 10.1124/jpet.118.255018.

Huang, R. S. *et al.* (2008) 'Identification of genetic variants and gene expression relationships associated with pharmacogenes in humans', *The Pharmacogenetics and Genomics Journal*, pp. 545–549.

Hulot, J.-S. *et al.* (2006) 'Cytochrome P450 2C19 loss-of-function polymorphism is a major determinant of clopidogrel responsiveness in healthy subjects', *Blood*, 108(7), pp. 2244–2247. doi: https://doi.org/10.1182/blood-2006-04-013052.

Hulot, J.-S., Hajjar, R. and Montalescot, G. (2012) 'Clopidogrel and CYP2C19 Testing: Ready for Clinical Prime Time?', *Clinical Chemistry*, 58(1), pp. 154–157. doi: 10.1373/clinchem.2011.165829.

Hunt, K. A. *et al.* (2008) 'Newly identified genetic risk variants for celiac disease related to the immune response', *Nature genetics*. 2008/03/02, 40(4), pp. 395–402. doi: 10.1038/ng.102.

Hustert, E. *et al.* (2001) 'The genetic determinants of the CYP3A5 polymorphism', *Pharmacogenetics and Genomics*, 11(9). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2001/12000/The_genetic_determinants_of_the_CYP3A5.5.aspx.

Ikeya, K. *et al.* (1989) 'Human CYP1A2: Sequence, Gene Structure, Comparison with the Mouse and Rat Orthologous Gene, and Differences in Liver 1A2 mRNA Expression', *Molecular Endocrinology*, 3(9), pp. 1399–1408. doi: 10.1210/mend-3-9-1399.

Imamura, C. K. *et al.* (2016) 'Impact of cytochrome P450 2C19 polymorphisms on the pharmacokinetics of tacrolimus when coadministered with voriconazole', *Journal of clinical pharmacology*. 2015/10/08, 56(4), pp. 408–413. doi: 10.1002/jcph.605.

Ingelman-Sundberg, M. *et al.* (2007) 'Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects', *Pharmacology & Therapeutics*, 116(3), pp. 496–526. doi: https://doi.org/10.1016/j.pharmthera.2007.09.004.

Innocenti, F. *et al.* (2011) 'Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue', *PLoS genetics*. 2011/05/26, 7(5), pp. e1002078–e1002078. doi: 10.1371/journal.pgen.1002078.

Iskakova, A. N. *et al.* (2016) 'Polymorphisms in genes involved in the absorption, distribution, metabolism, and excretion of drugs in the Kazakhs of Kazakhstan', *BMC Genetics*, 17(1). doi: 10.1186/s12863-016-0329-x.

Issa, A. M. (2002) 'Ethical perspectives on pharmacogenomic profiling in the drug development process', *Nature Reviews Drug Discovery*, 1(4), pp. 300–308. doi: 10.1038/nrd771.

Issa, A. M. (2007) 'Personalized medicine and the practice of medicine in the 21st century', *McGill journal of medicine : MJM : an international forum for the advancement of medical sciences by students*, 10(1), pp. 53–57. Available at: https://pubmed.ncbi.nlm.nih.gov/18523593.

Jacobs, B. A. W. *et al.* (2019) 'Pharmacokinetics of Capecitabine and Four Metabolites in a Heterogeneous Population of Cancer Patients: A Comprehensive Analysis', *CPT: pharmacometrics & systems pharmacology*. 2019/11/20, 8(12), pp. 940–950. doi: 10.1002/psp4.12474.

Jada, S. R. *et al.* (2007) 'Role of UGT1A1*6, UGT1A1*28 and ABCG2 c.421C>A polymorphisms in irinotecan-induced neutropenia in Asian cancer patients', *Cancer Science*, 98(9), pp. 1461–1467. doi: 10.1111/j.1349-7006.2007.00541.x.

Jakobsson, M., Edge, M. D. and Rosenberg, N. A. (2013) 'The Relationship Between FST and the Frequency of the Most Frequent Allele', *Genetics*, 193(2), pp. 515–528. doi: 10.1534/genetics.112.144758.

Jakoby, W. B. and Ziegler, D. M. (1990) 'The enzymes of detoxication.', *Journal of Biological Chemistry* , 265(34), pp. 20715–20718. Available at: http://www.jbc.org/content/265/34/20715.short.

Jancova, P., Anzenbacher, P. and Anzenbacherova, E. (2010) 'Phase II drug metabolizing enzymes', *Biomedical Papers*, 154(2), pp. 103–116. doi: 10.5507/bp.2010.017.

Jansen, R. C. and Nap, J.-P. (2001) 'Genetical genomics: the added value from segregation', *Trends in Genetics*, 17(7), pp. 388–391. doi: https://doi.org/10.1016/S0168-9525(01)02310-1.

Jennings, B. A. *et al.* (2013) 'Evaluating predictive pharmacogenetic signatures of adverse events in colorectal cancer patients treated with fluoropyrimidines', *PloS one*, 8(10), pp. e78053–e78053. doi: 10.1371/journal.pone.0078053.

Johnson, D. H. *et al.* (2014) 'Genomewide association study of atazanavir pharmacokinetics and hyperbilirubinemia in AIDS Clinical Trials Group protocol A5202', *Pharmacogenetics and genomics*, 24(4), pp. 195–203. doi: 10.1097/FPC.0000000000000034.

Johnson, M. R., Wang, K. and Diasio, R. B. (2002) 'Profound Dihydropyrimidine Dehydrogenase Deficiency Resulting from a Novel Compound Heterozygote Genotype', *Clinical Cancer Research*, 8(3), pp. 768 LP – 774. Available at: http://clincancerres.aacrjournals.org/content/8/3/768.abstract.

Johnston, H. R., Keats, B. J. B. and Sherman, S. L. (2019) *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition)*.

Jose, M. *et al.* (2016) 'Role of CYP2C19 gene polymorphism in acute alcohol withdrawal treatment with loading dose of diazepam in a South Indian population', *European Journal of Clinical Pharmacology*, 72(7), pp. 807–812. doi: 10.1007/s00228-016-2061-x.

Jung, S. *et al.* (2020) 'Expression Quantitative Trait Loci (eQTL) Mapping in Korean Patients With Crohn's Disease and Identification of Potential Causal Genes Through Integration With Disease Associations ', *Frontiers in Genetics* , p. 486. Available at: https://www.frontiersin.org/article/10.3389/fgene.2020.00486.

Kalacas, N. A. *et al.* (2019) 'GSTM1 and GSTT1 Genetic Polymorphisms and Breast Cancer

Risk in Selected Filipino Cases', *Asian Pacific journal of cancer prevention : APJCP*, 20(2), pp. 529–535. doi: 10.31557/APJCP.2019.20.2.529.

Kall, M. A. and Clausen, J. (1995) 'Dietary effect on mixed function P450 1A2 activity assayed by estimation of caffeine metabolism in man', *Human & Experimental Toxicology*, 14(10), pp. 801–807. doi: 10.1177/096032719501401004.

Kalow, W. and Gunn, D. R. (1957) 'THE RELATION BETWEEN DOSE OF SUCCINYLCHOLINE AND DURATION OF APNEA IN MAN', *Journal of Pharmacology and Experimental Therapeutics*, 120(2), pp. 203 LP – 214. Available at: http://jpet.aspetjournals.org/content/120/2/203.abstract.

Kaskas, B. A. *et al.* (2003) 'Safe treatment of thiopurine S-methyltransferase deficient Crohn's disease patients with azathioprine', *Gut*, 52(1), pp. 140–142. doi: 10.1136/gut.52.1.140.

Kishi, S. *et al.* (2004) 'Effects of prednisone and genetic polymorphisms on etoposide disposition in children with acute lymphoblastic leukemia', *Blood*, 103(1), pp. 67–72. doi: 10.1182/blood-2003-06-2105.

Kivistö, K. T. *et al.* (2004) 'Lipid-lowering response to statins is affected by CYP3A5 polymorphism', *Pharmacogenetics and Genomics*, 14(8). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2004/08000/Lipid_lowering_response_to_statins_is_affected_by.5.aspx.

Krishnakumar, D. *et al.* (2012) 'Genetic polymorphisms of drug-metabolizing phase I enzymes CYP2E1, CYP2A6 and CYP3A5 in South Indian population', *Fundamental & Clinical Pharmacology*, 26(2), pp. 295–306. doi: 10.1111/j.1472-8206.2010.00917.x.

Kudzi, W. *et al.* (2011) 'Pharmacogenetics in Ghana: reviewing the evidence', *Ghana medical journal*, 45(2), pp. 73–80. doi: 10.4314/gmj.v45i2.68932.

Kudzi, W. *et al.* (2016) 'Genetic polymorphisms of patients on stable warfarin maintenance therapy in a Ghanaian population', *BMC research notes*, 9(1), p. 507. doi: 10.1186/s13104-016-2306-x.

Kuehl, P. *et al.* (2001) 'Sequence diversity in CYP3A promoters and characterization of the

genetic basis of polymorphic CYP3A5 expression', *Nature Genetics*, 27(4), pp. 383–391. doi: 10.1038/86882.

Kuilenburg, A. B. P. van *et al.* (2016) 'Phenotypic and clinical implications of variants in the dihydropyrimidine dehydrogenase gene', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1862(4), pp. 754–762. doi: https://doi.org/10.1016/j.bbadis.2016.01.009.

Kuo, C.-H. *et al.* (2014) 'CYP2C19 polymorphism influences Helicobacter pylori eradication', *World journal of gastroenterology*, 20(43), pp. 16029–16036. doi: 10.3748/wjg.v20.i43.16029.

Kurth, J. H. (2000) 'Pharmacogenomics: Future Promise of a Tool for Identifying Patients at Risk', *Drug information journal : DIJ / Drug Information Association* , 34(1), pp. 223–227. doi: 10.1177/009286150003400130.

Laika, B. *et al.* (2010) 'Pharmacogenetics and olanzapine treatment: CYP1A2*1F and serotonergic polymorphisms influence therapeutic outcome', *The Pharmacogenomics Journal*, 10(1), pp. 20–29. doi: 10.1038/tpj.2009.32.

Lamba, J. *et al.* (2012) 'PharmGKB summary: very important pharmacogene information for CYP3A5', *Pharmacogenetics and genomics*, 22(7), pp. 555–558. doi: 10.1097/FPC.0b013e328351d47f.

Lankisch, T. O. *et al.* (2006) 'Gilbert's disease and atazanavir: From phenotype to UDP-glucuronosyltransferase haplotype', *Hepatology*, 44(5), pp. 1324–1332. doi: 10.1002/hep.21361.

Lankisch, T. O. *et al.* (2008) 'Gilbert's Syndrome and Irinotecan Toxicity: Combination with UDP-Glucuronosyltransferase 1A7 Variants Increases Risk', *Cancer Epidemiology Biomarkers & Prevention*, 17(3), pp. 695 LP – 701. doi: 10.1158/1055-9965.EPI-07-2517.

Largillier, R. *et al.* (2006) 'Pharmacogenetics of Capecitabine in Advanced Breast Cancer Patients', *Clinical Cancer Research*, 12(18), pp. 5496 LP – 5502. doi: 10.1158/1078-0432.CCR-06-0320.

Larson, A. M. *et al.* (2005) 'Acetaminophen-induced acute liver failure: Results of a United

States multicenter, prospective study', *Hepatology*, 42(6), pp. 1364–1372. doi: 10.1002/hep.20948.

Lauschke, V. M., Milani, L. and Ingelman-Sundberg, M. (2017) 'Pharmacogenomic Biomarkers for Improved Drug Therapy—Recent Progress and Future Developments', *The AAPS Journal*, 20(1), p. 4. doi: 10.1208/s12248-017-0161-x.

Lee, A. M. *et al.* (2014) 'DPYD Variants as Predictors of 5-fluorouracil Toxicity in Adjuvant Colon Cancer Treatment (NCCTG N0147)', *JNCI: Journal of the National Cancer Institute*, 106(12). doi: 10.1093/jnci/dju298.

Lee, C. R., Goldstein, J. A. and Pieper, J. A. (2002) 'Cytochrome P450 2C9 polymorphisms: a comprehensive review of the in-vitro and human data', *Pharmacogenetics and Genomics*, 12(3). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2002/04000/Cytochrome_P450_2C9_polymorphisms__a_comprehensive.10.aspx.

Lee, H.-K. *et al.* (2013) 'Effects of polymorphisms in ABCG2, SLCO1B1, SLC10A1 and CYP2C9/19 on plasma concentrations of rosuvastatin and lipid response in Chinese patients', *Pharmacogenomics*, 14(11), p. 1283—1294. doi: 10.2217/pgs.13.115.

Lee, S.-J. *et al.* (2003) 'Genetic findings and functional studies of human CYP3A5 single nucleotide polymorphisms in different ethnic groups*', *Pharmacogenetics and Genomics*, 13(8). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2003/08000/Genetic_findings_and_functional_studies_of_human.4.aspx.

Lepik, K. J. *et al.* (2018) 'Adverse drug reactions to integrase strand transfer inhibitors', *AIDS*, 32(7). Available at: https://journals.lww.com/aidsonline/Fulltext/2018/04240/Adverse_drug_reactions_to_integrase_strand.9.aspx.

Li-Wan-Po, A. *et al.* (2010) 'Pharmacogenetics of CYP2C19: Functional and clinical implications of a new variant CYP2C19*17', *British Journal of Clinical Pharmacology*, 69(3), pp. 222–230. doi: 10.1111/j.1365-2125.2009.03578.x.

Li, J. *et al.* (2011) 'Global patterns of genetic diversity and signals of natural selection for

human ADME genes', 20(3), pp. 528–540. doi: 10.1093/hmg/ddq498.

Li, J. *et al.* (2014) 'Increased genetic diversity of ADME genes in African Americans compared with their putative ancestral source populations and implications for Pharmacogenomics', *BMC Genetics*, 15, pp. 1–15. doi: 10.1186/1471-2156-15-52.

Libioulle, C. *et al.* (2007) 'Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4', *PLoS genetics*. 2007/03/05, 3(4), pp. e58–e58. doi: 10.1371/journal.pgen.0030058.

Limdi, N. A. *et al.* (2010) 'Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups', *Blood*. 2010/03/04, 115(18), pp. 3827–3834. doi: 10.1182/blood-2009-12-255992.

Limdi, N. A. and Veenstra, D. L. (2008) 'Warfarin Pharmacogenetics', *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 28(9), pp. 1084–1097. doi: 10.1592/phco.28.9.1084.

Lin, J. H. (2007) 'Pharmacokinetic and Pharmacodynamic Variability: A Daunting Challenge in Drug Therapy', *Current Drug Metabolism*, pp. 109–136. doi: http://dx.doi.org/10.2174/138920007779816002.

Lin, K.-M. *et al.* (2010) 'CYP1A2 genetic polymorphisms are associated with treatment response to the antidepressant paroxetine', *Pharmacogenomics*, 11(11), pp. 1535–1543. doi: 10.2217/pgs.10.128.

Liu, X. *et al.* (2014) 'Association of UGT1A1*28 polymorphisms with irinotecan-induced toxicities in colorectal cancer: a meta-analysis in Caucasians', *The pharmacogenomics journal*. 2013/03/26, 14(2), pp. 120–129. doi: 10.1038/tpj.2013.10.

Llibre, J. M. *et al.* (2019) 'Discontinuation of dolutegravir, elvitegravir/cobicistat and raltegravir because of toxicity in a prospective cohort', *HIV Medicine*, 20(3), pp. 237–247. doi: https://doi.org/10.1111/hiv.12710.

Loganayagam, A. *et al.* (2013) 'Pharmacogenetic variants in the DPYD, TYMS, CDA and MTHFR genes are clinically significant predictors of fluoropyrimidine toxicity', *British journal of cancer*. 2013/06/04, 108(12), pp. 2505–2515. doi: 10.1038/bjc.2013.262.

Lonsdale, J. *et al.* (2013) 'The Genotype-Tissue Expression (GTEx) project', *Nature Genetics*, 45(6), pp. 580–585. doi: 10.1038/ng.2653.

Lu, Y. *et al.* (2014) 'CYP3A5 genotype impacts maraviroc concentrations in healthy volunteers', *Drug metabolism and disposition: the biological fate of chemicals*. 2014/08/12, 42(11), pp. 1796–1802. doi: 10.1124/dmd.114.060194.

Luzzatto, L., Nannelli, C. and Notaro, R. (2016) 'Glucose-6-Phosphate Dehydrogenase Deficiency', *Hematology/Oncology Clinics of North America*, 30(2), pp. 373–393. doi: https://doi.org/10.1016/j.hoc.2015.11.006.

Ma, Q. and Lu, A. Y. H. (2011) 'Pharmacogenetics, Pharmacogenomics, and Individualized Medicine', *Pharmacological Reviews*. Edited by D. R. Sibley, 63(2), pp. 437 LP – 459. doi: 10.1124/pr.110.003533.

MacPhee, I. A. M. *et al.* (2004) 'The Influence of Pharmacogenetics on the Time to Achieve Target Tacrolimus Concentrations after Kidney Transplantation', *American Journal of Transplantation*, 4(6), pp. 914–919. doi: 10.1111/j.1600-6143.2004.00435.x.

Manolio, T. A. *et al.* (2009) 'Finding the missing heritability of complex diseases', *Nature*, 461(7265), pp. 747–753. doi: 10.1038/nature08494.

Marian, A. J. (2009) 'Cytochrome p-450 polymorphisms and response to clopidogrel.', *Current atherosclerosis reports*, 11(3), pp. 157–160. doi: 10.1007/s11883-009-0025-7.

Marshall, A. (1997) 'Getting the right drug into the right patient', *Nature Biotechnology*, 15(12), pp. 1249–1252. doi: 10.1038/nbt1197-1249.

Mason, E. and Russell, D. W. (1971) 'Isoniazid acetylation rates (phenotypes) of patients being treated for tuberculosis', *Bulletin of the World Health Organization*, 45(5), pp. 617–624. Available at: https://pubmed.ncbi.nlm.nih.gov/5316954.

Mathijssen, R. H. J. *et al.* (2001) 'Clinical Pharmacokinetics and Metabolism of Irinotecan (CPT-11)', *Clinical Cancer Research*, 7(8), pp. 2182 LP – 2194. Available at: http://clincancerres.aacrjournals.org/content/7/8/2182.abstract.

McCune, J. S. *et al.* (2005) 'CONTRIBUTION OF CYP3A5 TO HEPATIC AND RENAL IFOSFAMIDE N-DECHLOROETHYLATION', *Drug Metabolism and Disposition*, 33(7),

pp. 1074 LP – 1081. doi: 10.1124/dmd.104.002279.

McDonagh, E. M. *et al.* (2012) 'PharmGKB summary: very important pharmacogene information for G6PD', *Pharmacogenetics and genomics*, 22(3), pp. 219–228. doi: 10.1097/FPC.0b013e32834eb313.

McGill, M. R. and Jaeschke, H. (2013) 'Metabolism and disposition of acetaminophen: recent advances in relation to hepatotoxicity and diagnosis', *Pharmaceutical research*. 2013/03/06, 30(9), pp. 2174–2187. doi: 10.1007/s11095-013-1007-6.

McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), p. 122. doi: 10.1186/s13059-016-0974-4.

Medeiros, R. *et al.* (2003) 'Platinum/paclitaxel-based chemotherapy in advanced ovarian carcinoma: glutathione S-transferase genetic polymorphisms as predictive biomarkers of disease outcome', *International Journal of Clinical Oncology*, 8(3), pp. 156–161. doi: 10.1007/s10147-003-0318-8.

Mercadel, C. J. *et al.* (2014) 'Dolutegravir: An Integrase Strand Transfer Inhibitor for the Treatment of Human Immunodeficiency Virus 1 in Adults', *The Journal of pharmacy technology : jPT : official publication of the Association of Pharmacy Technicians*. 2014/07/22, 30(6), pp. 216–226. doi: 10.1177/8755122514544126.

Meulendijks, D. *et al.* (2015) 'Clinical relevance of DPYD variants c.1679T>G, c.1236G>A/HapB3, and c.1601G>A as predictors of severe fluoropyrimidine-associated toxicity: a systematic review and meta-analysis of individual patient data', *The Lancet Oncology*, 16(16), pp. 1639–1650. doi: 10.1016/S1470-2045(15)00286-7.

Meyer, U. A. (2000) 'Pharmacogenetics and adverse drug reactions', *The Lancet*, 356(9242), pp. 1667–1671. doi: 10.1016/S0140-6736(00)03167-6.

Meyer, U. A., Zanger, U. M. and Schwab, M. (2013) 'Omics and Drug Response', *Annual Review of Pharmacology and Toxicology*, 53(1), pp. 475–502. doi: 10.1146/annurev-pharmtox-010510-100502.

Milne, R. L. *et al.* (2017) 'Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer', *Nature genetics*. 2017/10/23, 49(12), pp. 1767–1778. doi:

10.1038/ng.3785.

Mirabbasi, S. A. *et al.* (2017) 'CYP2C19 genetic variation and individualized clopidogrel prescription in a cardiology clinic', *Journal of community hospital internal medicine perspectives*, 7(3), pp. 151–156. doi: 10.1080/20009666.2017.1347475.

Moffatt, M. F. *et al.* (2007) 'Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma', *Nature*, 448(7152), pp. 470–473. doi: 10.1038/nature06014.

Moorhouse, M. A. *et al.* (2018) 'Southern African HIV Clinicians Society Guidance on the use of dolutegravir in first-line antiretroviral therapy', *Southern African Journal of HIV Medicine; Vol 19, No 1 (2018)DO - 10.4102/sajhivmed.v19i1.917* . Available at: https://sajhivmed.org.za/index.php/hivmed/article/view/917/1276.

Morel, A. *et al.* (2006) 'Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide polymorphisms on 5-fluorouracil tolerance', *Molecular Cancer Therapeutics*, 5(11), pp. 2895 LP – 2904. doi: 10.1158/1535-7163.MCT-06-0327.

Mouly, S. J. *et al.* (2005) 'Variation in oral clearance of saquinavir is predicted by CYP3A5*1 genotype but not by enterocyte content of cytochrome P450 3A5', *Clinical Pharmacology & Therapeutics*, 78(6), pp. 605–618. doi: 10.1016/j.clpt.2005.08.014.

Mrazek, D. A. *et al.* (2011) 'CYP2C19 variation and citalopram response', *Pharmacogenetics and genomics*, 21(1), pp. 1–9. doi: 10.1097/fpc.0b013e328340bc5a.

Nica, A. C. and Dermitzakis, E. T. (2008) 'Using gene expression to investigate the genetic basis of complex disorders', *Human molecular genetics*, 17(R2), pp. R129–R134. doi: 10.1093/hmg/ddn285.

Nica, A. C. and Dermitzakis, E. T. (2013) 'Expression quantitative trait loci: Present and future', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620). doi: 10.1098/rstb.2012.0362.

Nicolae, D. L. *et al.* (2010) 'Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS', *PLOS Genetics*, 6(4), p. e1000888. Available at: https://doi.org/10.1371/journal.pgen.1000888.

Ning, B. *et al.* (2005) 'Common genetic polymorphisms in the 5'-flanking Region of the

SULT1A1 gene: haplotypes and their association with platelet enzymatic activity', *Pharmacogenetics and Genomics*, 15(7). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2005/07000/Common_genetic_polymorphisms_in_the_5__flanking.4.aspx.

Nishijima, T. *et al.* (2014) 'Single-nucleotide polymorphisms in the UDP-glucuronosyltransferase 1A-3' untranslated region are associated with atazanavir-induced nephrolithiasis in patients with HIV-1 infection: a pharmacogenetic study', *Journal of Antimicrobial Chemotherapy*, 69(12), pp. 3320–3328. doi: 10.1093/jac/dku304.

O'Brien, H. E. *et al.* (2018) 'Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders', *Genome Biology*, 19(1), p. 194. doi: 10.1186/s13059-018-1567-1.

O'Seaghdha, C. M. *et al.* (2009) 'Higher tacrolimus trough levels on days 2–5 post-renal transplant are associated with reduced rates of acute rejection', *Clinical Transplantation*, 23(4), pp. 462–468. doi: 10.1111/j.1399-0012.2009.01021.x.

O'Donnell, P. H. and Dolan, M. E. (2009) 'Cancer Pharmacoethnicity: Ethnic Differences in Susceptibility to the Effects of Chemotherapy', *Clinical Cancer Research*, 15(15), pp. 4806 LP – 4814. doi: 10.1158/1078-0432.CCR-09-0344.

Obradovic, M., Mrhar, A. and Kos, M. (2008) 'Cost–effectiveness of UGT1A1 genotyping in second-line, high-dose, once every 3 weeks irinotecan monotherapy treatment of colorectal cancer', *Pharmacogenomics*, 9(5), pp. 539–549. doi: 10.2217/14622416.9.5.539.

Offer, S. M. *et al.* (2014) 'Comparative Functional Analysis of DPYD Variants of Potential Clinical Relevance to Dihydropyrimidine Dehydrogenase Activity', *Cancer Research*, 74(9), pp. 2545 LP – 2554. doi: 10.1158/0008-5472.CAN-13-2482.

Ohlsson Rosenborg, S. *et al.* (2008) 'Kinetics of omeprazole and escitalopram in relation to the CYP2C19*17 allele in healthy subjects', *European Journal of Clinical Pharmacology*, 64(12), pp. 1175–1179. doi: 10.1007/s00228-008-0529-z.

Onoue, M. *et al.* (2009) 'UGT1A1*6 polymorphism is most predictive of severe neutropenia induced by irinotecan in Japanese cancer patients', *International Journal of Clinical Oncology*, 14(2), pp. 136–142. doi: 10.1007/s10147-008-0821-z.

Ortega, V. E. and Meyers, D. A. (2014) 'Pharmacogenetics: Implications of race and ethnicity on defining genetic profiles for personalized medicine', *Journal of Allergy and Clinical Immunology*, 133(1), pp. 16–26. doi: 10.1016/j.jaci.2013.10.040.

Özdemir, V. *et al.* (2001) 'Treatment-Resistance to Clozapine in Association With Ultrarapid CYP1A2 Activity and the C → A Polymorphism in Intron 1 of the CYP1A2 Gene: Effect of Grapefruit Juice and Low-Dose Fluvoxamine', *Journal of Clinical Psychopharmacology*, 21(6). Available at: https://journals.lww.com/psychopharmacology/Fulltext/2001/12000/Treatment_Resistance_to_Clozapine_in_Association.11.aspx.

Palomaki, G. E. *et al.* (2009) 'Can UGT1A1 genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with irinotecan? An evidence-based review', *Genetics in medicine : official journal of the American College of Medical Genetics*, 11(1), pp. 21–34. doi: 10.1097/GIM.0b013e31818efd77.

Park, J. Y., Cha, Y. J. and Kim, K. A. (2014) 'CYP3A5*3 polymorphism and its clinical implications and pharmacokinetic role', *Translational and Clinical Pharmacology*, 22(1), pp. 03–07. doi: 10.12793/tcp.2014.22.1.3.

Park, S. Y. *et al.* (2008) 'Frequencies of CYP3A5 genotypes and haplotypes in a Korean population', *Journal of Clinical Pharmacy and Therapeutics*, 33(1), pp. 61–65. doi: 10.1111/j.1365-2710.2008.00879.x.

Passey, C. *et al.* (2011) 'Dosing equation for tacrolimus using genetic variants and clinical factors', *British Journal of Clinical Pharmacology*, 72(6), pp. 948–957. doi: 10.1111/j.1365-2125.2011.04039.x.

Patterson, A. D., Gonzalez, F. J. and Idle, J. R. (2010) 'Xenobiotic metabolism: a view through the metabolometer', *Chemical research in toxicology*, 23(5), pp. 851–860. doi: 10.1021/tx100020p.

Peacock, M. L. *et al.* (1993) 'Novel polymorphism in the A4 region of the amyloid precursor protein gene in a patient without Alzheimer's disease', *Neurology*, 43(6), pp. 1254 LP – 1254. doi: 10.1212/WNL.43.6.1254.

Peng, B. *et al.* (2004) 'Pharmacokinetics and Pharmacodynamics of Imatinib in a Phase I Trial

With Chronic Myeloid Leukemia Patients', *Journal of Clinical Oncology*, 22(5), pp. 935–942. doi: 10.1200/JCO.2004.03.050.

Peng, B., Lloyd, P. and Schran, H. (2005) 'Clinical Pharmacokinetics of Imatinib', *Clinical Pharmacokinetics*, 44(9), pp. 879–894. doi: 10.2165/00003088-200544090-00001.

Petrova, D. T. *et al.* (2007) 'Genotyping of CYP3A5 Polymorphisms among Bulgarian Patients with Sporadic Colorectal Cancer and Controls', *Oncology Research and Treatment*, 30(11), pp. 559–563. doi: 10.1159/000108285.

Phan, V. H. *et al.* (2009) 'Ethnic differences in drug metabolism and toxicity from chemotherapy', *Expert Opinion on Drug Metabolism & Toxicology*, 5(3), pp. 243–257. doi: 10.1517/17425250902800153.

Pirmohamed, M. *et al.* (2004) 'Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients', *BMJ (Clinical research ed.)*, 329(7456), pp. 15–19. doi: 10.1136/bmj.329.7456.15.

Playing, R., Croom, E. and Science, T. (2014) 'Xenobiotic Learn more about Xenobiotic Toxicology and Human Environments Toxicokinetics and Toxicodynamics Hormones of the Limbic System'.

Preissner, S. *et al.* (2015) 'Chapter Four - Personalized Cancer Therapy Considering Cytochrome P450 Variability', in Hardwick, J. P. B. T.-A. in P. (ed.) *Cytochrome P450 Function and Pharmacological Roles in Inflammation and Cancer*. Academic Press, pp. 113–130. doi: https://doi.org/10.1016/bs.apha.2015.03.004.

Price, A. L. *et al.* (2008) 'Effects of cis and trans genetic ancestry on gene expression in African Americans', *PLoS genetics*. 2008/12/05, 4(12), pp. e1000294–e1000294. doi: 10.1371/journal.pgen.1000294.

Quteineh, L. *et al.* (2008) 'Influence of CYP3A5 Genetic Polymorphism on Tacrolimus Daily Dose Requirements and Acute Rejection in Renal Graft Recipients', *Basic & Clinical Pharmacology & Toxicology*, 103(6), pp. 546–552. doi: 10.1111/j.1742-7843.2008.00327.x.

Raftogianis, R. B. *et al.* (1997) 'Phenol Sulfotransferase Pharmacogenetics in Humans: Association of CommonSULT1A1Alleles with TS PST Phenotype', *Biochemical and Biophysical*

Research Communications, 239(1), pp. 298–304. doi: https://doi.org/10.1006/bbrc.1997.7466.

Raida, M. *et al.* (2001) 'Prevalence of a Common Point Mutation in the Dihydropyrimidine Dehydrogenase (DPD) Gene within the 5'-Splice Donor Site of Intron 14 in Patients with Sever', *Clinical Cancer Research*, 7(9), pp. 2832 LP – 2839. Available at: http://clincancerres.aacrjournals.org/content/7/9/2832.abstract.

Ramesh, M., Ahlawat, P. and Srinivas, N. R. (2010) 'Irinotecan and its active metabolite, SN-38: review of bioanalytical methods and recent update from clinical pharmacology perspectives', *Biomedical Chromatography*, 24(1), pp. 104–123. doi: 10.1002/bmc.1345.

Ramos, E. *et al.* (2014) 'Pharmacogenomics, ancestry and clinical decision making for global populations', *Pharmacogenomics Journal*, 14(3), pp. 217–222. doi: 10.1038/tpj.2013.24.

Rasool, M. and Liu, M.-C. (2019) 'Effects of the Human SULT1A1 Polymorphisms on the Sulfation of Acetaminophen, O-Desmethylnaproxen, and Tapentadol', *The FASEB Journal*, 33(S1), pp. 473.7-473.7. doi: 10.1096/fasebj.2019.33.1_supplement.473.7.

Raucy, J. L. *et al.* (2002) 'Expression and Induction of CYP2C P450 Enzymes in Primary Cultures of Human Hepatocytes', *Journal of Pharmacology and Experimental Therapeutics*, 302(2), pp. 475 LP – 482. doi: 10.1124/jpet.102.033837.

Relling, M. V *et al.* (1992) 'Racial and gender differences in N-acetyltransferase, xanthine oxidase, and CYP1A2* activities', *Clinical Pharmacology & Therapeutics*, 52(6), pp. 643–658. doi: 10.1038/clpt.1992.203.

Renders, L. *et al.* (2007) 'CYP3A5 Genotype Markedly Influences the Pharmacokinetics of Tacrolimus and Sirolimus in Kidney Transplant Recipients', *Clinical Pharmacology & Therapeutics*, 81(2), pp. 228–234. doi: 10.1038/sj.clpt.6100039.

Richardson, M. *et al.* (2018) 'CYP genetic variants and toxicity related to anti-tubercular agents: a systematic review and meta-analysis', *Systematic reviews*, 7(1), p. 204. doi: 10.1186/s13643-018-0861-z.

Rieder, M. J. *et al.* (2005) 'Effect of VKORC1 Haplotypes on Transcriptional Regulation and

Warfarin Dose', *New England Journal of Medicine*, 352(22), pp. 2285–2293. doi: 10.1056/NEJMoa044503.

Robinson, M. and Horn, J. (2003) 'Clinical Pharmacology of Proton Pump Inhibitors', *Drugs*, 63(24), pp. 2739–2754. doi: 10.2165/00003495-200363240-00004.

Roden, D. M. *et al.* (2011) 'Pharmacogenomics: the genetics of variable drug responses', *Circulation*, 123(15), pp. 1661–1670. doi: 10.1161/CIRCULATIONAHA.109.914820.

Rodrigues, J. C. G. *et al.* (2019) 'Polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians', *Scientific Reports*, 9(1), p. 7201. doi: 10.1038/s41598-019-43610-y.

Roses, A. D. (2000) 'Pharmacogenetics and the practice of medicine', *Nature*, 405(6788), pp. 857–865. doi: 10.1038/35015728.

Rosmarin, D. *et al.* (2015) 'A candidate gene study of capecitabine-related toxicity in colorectal cancer identifies new toxicity variants at DPYD and a putative role for ENOSF1 rather than TYMS', *Gut*. 2014/03/19, 64(1), pp. 111–120. doi: 10.1136/gutjnl-2013-306571.

Rotimi, C. N. *et al.* (2016) 'The African diaspora: history, adaptation and health', *Current opinion in genetics & development*. 2016/09/16, 41, pp. 77–84. doi: 10.1016/j.gde.2016.08.005.

El Rouby, N., Lima, J. J. and Johnson, J. A. (2018) 'Proton pump inhibitors: from CYP2C19 pharmacogenetics to precision medicine', *Expert opinion on drug metabolism & toxicology*. 2018/04/12, 14(4), pp. 447–460. doi: 10.1080/17425255.2018.1461835.

Roy, J.-N. *et al.* (2005) 'CYP3A5 GENETIC POLYMORPHISMS IN DIFFERENT ETHNIC POPULATIONS', *Drug Metabolism and Disposition*, 33(7), pp. 884 LP – 887. doi: 10.1124/dmd.105.003822.

Rudberg, I. *et al.* (2008) 'Impact of the Ultrarapid CYP2C19*17 Allele on Serum Concentration of Escitalopram in Psychiatric Patients', *Clinical Pharmacology & Therapeutics*, 83(2), pp. 322–327. doi: 10.1038/sj.clpt.6100291.

SAIF, M. W. (2013) 'Dihydropyrimidine Dehydrogenase Gene (DPYD) Polymorphism among Caucasian and non-Caucasian Patients with 5-FU- and Capecitabine-related Toxicity Using

Full Sequencing of DPYD', *Cancer Genomics - Proteomics* , 10(2), pp. 89–92. Available at: http://cgp.iiarjournals.org/content/10/2/89.abstract.

Saiz-Rodríguez, M. *et al.* (2020) 'Effect of the Most Relevant CYP3A4 and CYP3A5 Polymorphisms on the Pharmacokinetic Parameters of 10 CYP3A Substrates', *Biomedicines*, 8(4), p. 94. doi: 10.3390/biomedicines8040094.

Saldivar, J. S. *et al.* (2016) 'Initial assessment of the benefits of implementing pharmacogenetics into the medical management of patients in a long-term care facility', *Pharmacogenomics and Personalized Medicine*, 9, pp. 1–6. doi: 10.2147/PGPM.S93480.

Salgueiro, N. *et al.* (2004) 'Mutations in exon 14 of dihydropyrimidine dehydrogenase and 5-Fluorouracil toxicity in Portuguese colorectal cancer patients', *Genetics in Medicine*, 6(2), pp. 102–107. doi: 10.1097/01.GIM.0000118061.66602.A5.

Sangkuhl, K., Klein, T. E. and Altman, R. B. (2010) 'Clopidogrel pathway', *Pharmacogenetics and genomics*, 20(7), pp. 463–465. doi: 10.1097/FPC.0b013e3283385420.

Sangkuhl, K., Klein, T. E. and Altman, R. B. (2011) 'PharmGKB summary: citalopram pharmacokinetics pathway', *Pharmacogenetics and genomics*, 21(11), pp. 769–772. doi: 10.1097/FPC.0b013e328346063f.

Savci-Heijink, C. D. *et al.* (2019) 'A specific gene expression signature for visceral organ metastasis in breast cancer', *BMC Cancer*, 19(1), p. 333. doi: 10.1186/s12885-019-5554-z.

Savic, R. M. *et al.* (2012) 'Effect of adherence as measured by MEMS, ritonavir boosting, and CYP3A5 genotype on atazanavir pharmacokinetics in treatment-naive HIV-infected patients', *Clinical pharmacology and therapeutics*. 2012/10/03, 92(5), pp. 575–583. doi: 10.1038/clpt.2012.137.

Schadt, E. E. *et al.* (2008) 'Mapping the genetic architecture of gene expression in human liver', *PLoS biology*, 6(5), pp. e107–e107. doi: 10.1371/journal.pbio.0060107.

van Schaik, R. H. N. *et al.* (2002) 'CYP3A5 Variant Allele Frequencies in Dutch Caucasians', *Clinical Chemistry*, 48(10), pp. 1668–1671. doi: 10.1093/clinchem/48.10.1668.

van Schaik, R. H. N. (2008) '6. Dose Adjustments Based on Pharmacogenetics of CYP450 Enzymes.', *Ejifcc*, 19(1), pp. 42–7. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/27683289%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4975340.

Schelleman, H. *et al.* (2010) 'New genetic variant that might improve warfarin dose prediction in African Americans', *British journal of clinical pharmacology*, 70(3), pp. 393–399. doi: 10.1111/j.1365-2125.2010.03709.x.

Schröder, A. *et al.* (2013) 'Genomics of ADME gene expression: Mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver', *Pharmacogenomics Journal*, 13(1), pp. 12–20. doi: 10.1038/tpj.2011.44.

Schulz, C. *et al.* (2009) 'UGT1A1 genotyping: a predictor of irinotecan-associated side effects and drug efficacy?', *Anti-cancer drugs*, 20(10), p. 867—879. doi: 10.1097/cad.0b013e328330c7d2.

Schwab, M. *et al.* (2008) 'Role of genetic and nongenetic factors for fluorouracil treatment-related severe toxicity: A prospective clinical trial by the German 5-FU toxicity study group', *Journal of Clinical Oncology*, 26(13), pp. 2131–2138. doi: 10.1200/JCO.2006.10.4182.

Schwab, M. and Schaeffeler, E. (2011) 'Warfarin pharmacogenetics meets clinical use', *Blood*, 118(11), pp. 2938–2939. doi: 10.1182/blood-2011-07-367763.

Schwab, M. and Schaeffeler, E. (2012a) 'Pharmacogenomics: a key component of personalized therapy', *Genome Medicine*, 4(11). doi: 10.1186/gm394.

Schwab, M. and Schaeffeler, E. (2012b) 'Pharmacogenomics: a key component of personalized therapy', *Genome medicine*, 4(11), p. 93. doi: 10.1186/gm394.

Sconce, E. A. *et al.* (2005) 'The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen', *Blood*, 106(7), pp. 2329–2333. doi: 10.1182/blood-2005-03-1108.

Scott, S. A. *et al.* (2013) 'Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update', *Clinical pharmacology and therapeutics*. 2013/05/22, 94(3), pp. 317–323. doi: 10.1038/clpt.2013.105.

Sebastiani, P. *et al.* (2009) 'Genome-wide association studies and the genetic dissection of

complex traits', *American journal of hematology*, 84(8), pp. 504–515. doi: 10.1002/ajh.21440.

Seidman, E. G. (2003) 'Clinical use and practical application of TPMT enzyme and 6-mercaptopurine metabolite monitoring in IBD', *Reviews in Gastroenterological Disorders*, 3(SUPPL. 1).

Sengupta, D. *et al.* (2020) 'Genetic-substructure and complex demographic history of South African Bantu speakers', *bioRxiv*, p. 2020.08.11.243840. doi: 10.1101/2020.08.11.243840.

Shabalin, A. A. (2012) 'Matrix eQTL: Ultra fast eQTL analysis via large matrix operations', *Bioinformatics*, 28(10), pp. 1353–1358. doi: 10.1093/bioinformatics/bts163.

Shih, P.-S. and Huang, J.-D. (2002) 'Pharmacokinetics of Midazolam and 1'-Hydroxymidazolam in Chinese with Different CYP3A5 Genotypes', *Drug Metabolism and Disposition*, 30(12), pp. 1491 LP – 1496. doi: 10.1124/dmd.30.12.1491.

Shimada, T. *et al.* (1994) 'Interindividual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians.', *Journal of Pharmacology and Experimental Therapeutics*, 270(1), pp. 414 LP – 423. Available at: http://jpet.aspetjournals.org/content/270/1/414.abstract.

Shin, J. M. and Kim, N. (2013) 'Pharmacokinetics and pharmacodynamics of the proton pump inhibitors', *Journal of neurogastroenterology and motility*. 2013/01/08, 19(1), pp. 25–35. doi: 10.5056/jnm.2013.19.1.25.

Shriner, D. *et al.* (2014) 'Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry', *Scientific Reports*, 4(1), p. 6055. doi: 10.1038/srep06055.

Shu, W. *et al.* (2017) 'Cytochrome P450 Genetic Variations Can Predict mRNA Expression, Cyclophosphamide 4-Hydroxylation, and Treatment Outcomes in Chinese Patients With Non-Hodgkin's Lymphoma', *The Journal of Clinical Pharmacology*, 57(7), pp. 886–898. doi: 10.1002/jcph.878.

Sibbing, D. *et al.* (2010) 'Cytochrome 2c19*17 allelic variant, platelet aggregation, bleeding events, and stent thrombosis in clopidogrel-treated patients with coronary stent placement', *Circulation*, 121(4), pp. 512–518. doi: 10.1161/CIRCULATIONAHA.109.885194.

Sim, S. C. *et al.* (2006) 'A common novel CYP2C19 gene variant causes ultrarapid drug metabolism relevant for the drug response to proton pump inhibitors and antidepressants', *Clinical Pharmacology & Therapeutics*, 79(1), pp. 103–113. doi: 10.1016/j.clpt.2005.10.002.

Single, R. M. and Thomson, G. (2016) 'Linkage Disequilibrium: Population Genetics of Multiple Loci', in Kliman, R. M. B. T.-E. of E. B. (ed.). Oxford: Academic Press, pp. 400–404. doi: https://doi.org/10.1016/B978-0-12-800049-6.00030-5.

Sissung, T. M. *et al.* (2012) 'Transporter pharmacogenetics: transporter polymorphisms affect normal physiology, diseases, and pharmacotherapy', *Discovery medicine*, 13(68), pp. 19–34. Available at: https://pubmed.ncbi.nlm.nih.gov/22284781.

Skrzydło-Radomańska, B. and Radwan, P. (2015) 'Dexlansoprazole - a new-generation proton pump inhibitor', *Przeglad gastroenterologiczny*. 2015/12/16, 10(4), pp. 191–196. doi: 10.5114/pg.2015.56109.

Slatkin, M. (2008) 'Linkage disequilibrium — understanding the evolutionary past and mapping the medical future', *Nature Reviews Genetics*, 9(6), pp. 477–485. doi: 10.1038/nrg2361.

Smith, M. (2020) '9A - Personalized precision medicine—Part A: Concepts and relevance in Mendelian disorders', in Smith, M. B. T.-G. E. I. (ed.). Academic Press, pp. 227–257. doi: https://doi.org/10.1016/B978-0-12-819613-7.00009-8.

Sohn, D.-R. *et al.* (1997) 'Metabolic disposition of lansoprazole in relation to the S-mephenytoin 4'-hydroxylation phenotype status', *Clinical Pharmacology & Therapeutics*, 61(5), pp. 574–582. doi: 10.1016/S0009-9236(97)90137-5.

Spear, B. B., Heath-Chiozzi, M. and Huff, J. (2001) 'Clinical application of pharmacogenetics', *Trends in Molecular Medicine*, 7(5), pp. 201–204. doi: 10.1016/S1471-4914(01)01986-4.

Stingl, J. and Viviani, R. (2014) 'Polymorphism in CYP2D6 and CYP2C19 , members of the cytochrome P450 mixed-function oxidase system , in the metabolism of psychotropic drugs',

pp. 167–177. doi: 10.1111/joim.12317.

Stranger, B. E. *et al.* (2012) 'Patterns of Cis Regulatory Variation in Diverse Human Populations', *PLOS Genetics*, 8(4), p. e1002639. Available at: https://doi.org/10.1371/journal.pgen.1002639.

Strassburg, C. P. (2008) 'Pharmacogenetics of Gilbert's syndrome', *Pharmacogenomics*, 9(6), pp. 703–715. doi: 10.2217/14622416.9.6.703.

Stubbins, M. J. *et al.* (1996) 'Genetic analysis of the human cytochrome P450 CYP2C9 locus', *Pharmacogenetics and Genomics*, 6(5). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/1996/10000/Genetic_analysis_of_the_human_cytochrome_P450.7.aspx.

Suarez-Kurtz, G. *et al.* (2014) 'Global Pharmacogenomics: Distribution of CYP3A5 Polymorphisms and Phenotypes in the Brazilian Population', *PLOS ONE*, 9(1), p. e83472. Available at: https://doi.org/10.1371/journal.pone.0083472.

Sugarman, E. A. *et al.* (2016) 'Contribution of Pharmacogenetic Testing to Modeled Medication Change Recommendations in a Long-Term Care Population with Polypharmacy', *Drugs & aging*, 33(12), pp. 929–936. doi: 10.1007/s40266-016-0412-z.

Sukasem, C. *et al.* (2016) 'Clinically relevant genetic variants of drug-metabolizing enzyme and transporter genes detected in Thai children and adolescents with autism spectrum disorder', *Neuropsychiatric Disease and Treatment*, 12, p. 843. doi: 10.2147/NDT.S101580.

Sulzyc-Bielicka, V. *et al.* (2008) '5-Fluorouracil toxicity-attributable IVS14 + 1G > A mutation of the dihydropyrimidine dehydrogenase gene in Polish colorectal cancer patients', *Pharmacological Reports*, 60(2), pp. 238–242.

Sumida, K. *et al.* (2013) 'Importance of UDP-glucuronosyltransferase 1A1 expression in skin and its induction by UVB in neonatal hyperbilirubinemia', *Molecular pharmacology*. 2013/08/15, 84(5), pp. 679–686. doi: 10.1124/mol.113.088112.

Swen, J. J. *et al.* (2007) 'Translating Pharmacogenomics: Challenges on the Road to the Clinic', *PLOS Medicine*, 4(8), p. e209. Available at: https://doi.org/10.1371/journal.pmed.0040209.

Tabrizi, A. R. *et al.* (2002) 'The frequency and effects of cytochrome P450 (CYP) 2C9 polymorphisms in patients receiving warfarin[1] ', *Journal of the American College of Surgeons*, 194(3), pp. 267–273. doi: 10.1016/S1072-7515(01)01163-2.

Takahashi, H. *et al.* (2003) 'Population differences in S-warfarin metabolism between CYP2C9 genotype-matched Caucasian and Japanese patients', *Clinical Pharmacology & Therapeutics*, 73(3), pp. 253–263. doi: 10.1067/mcp.2003.26a.

Takahashi, H. *et al.* (2006) 'Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese, Caucasians and African-Americans', *Pharmacogenetics and Genomics*, 16(2). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2006/02000/Different_contributions_of_polymorphisms_in_VKORC1.4.aspx.

Takano, M. *et al.* (2009) 'Clinical Significance of UDP-Glucuronosyltransferase 1A1*6 for Toxicities of Combination Chemotherapy with Irinotecan and Cisplatin in Gynecologic Cancers', *Oncology*, 76(5), pp. 315–321. doi: 10.1159/000209335.

Takeuchi, F. *et al.* (2009) 'A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose', *PLoS genetics*. 2009/03/20, 5(3), pp. e1000433–e1000433. doi: 10.1371/journal.pgen.1000433.

Tam, V. C. *et al.* (2009) 'Generalizability of toxicity data from oncology clinical trials to clinical practice: toxicity of irinotecan-based regimens in patients with metastatic colorectal cancer', *Current oncology (Toronto, Ont.)*, 16(6), pp. 13–20. doi: 10.3747/co.v16i6.426.

Tang, H.-L. *et al.* (2013) 'Effects of CYP2C19 loss-of-function variants on the eradication of H. pylori infection in patients treated with proton pump inhibitor-based triple therapy regimens: a meta-analysis of randomized clinical trials', *PloS one*, 8(4), pp. e62162–e62162. doi: 10.1371/journal.pone.0062162.

Tangkhuenkhan, P. *et al.* (2020) 'Frequency and Association Of GSTM1 and GSTT1 Gene Polymorphisms with Survival in Breast Cancer Patients', *Asian Pacific journal of cancer prevention : APJCP*, 21(8), pp. 2251–2257. doi: 10.31557/APJCP.2020.21.8.2251.

Tay, J. K. X. *et al.* (2007) 'Functional polymorphisms of the cytochrome P450 1A2 (CYP1A2) gene and prolonged QTc interval in schizophrenia', *Progress in Neuro-Psychopharmacology*

and Biological Psychiatry, 31(6), pp. 1297–1302. doi: https://doi.org/10.1016/j.pnpbp.2007.05.013.

Tecza, K. *et al.* (2018) 'Pharmacogenetics of toxicity of 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer patients', *Oncotarget*, 9(10), pp. 9114–9136. doi: 10.18632/oncotarget.24148.

Teo, Y. Y., Small, K. S. and Kwiatkowski, D. P. (2010) 'Methodological challenges of genome-wide association analysis in Africa', *Nature Reviews Genetics*, 11(2), pp. 149–160. doi: 10.1038/nrg2731.

Terrazzino, S. *et al.* (2013) 'DPYD IVS14+1G>A and 2846A>T genotyping for the prediction of severe fluoropyrimidine-related toxicity: a meta-analysis', *Pharmacogenomics*, 14(11), pp. 1255–1272. doi: 10.2217/pgs.13.116.

Thami, P. K. and Chimusa, E. R. (2019) 'Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes Within Southern Africa ', *Frontiers in Genetics* , p. 905. Available at: https://www.frontiersin.org/article/10.3389/fgene.2019.00905.

Thomas, F. *et al.* (2016) 'Genotyping of a family with a novel deleterious DPYD mutation supports the pretherapeutic screening of DPD deficiency with dihydrouracil/uracil ratio', *Clinical Pharmacology & Therapeutics*, 99(2), pp. 235–242. doi: 10.1002/cpt.210.

Thorn, C. F. *et al.* (2012) 'PharmGKB summary: very important pharmacogene information for CYP1A2', *Pharmacogenetics and Genomics*, 22(1). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2012/01000/PharmGKB_summary__very_important_pharmacogene.9.aspx.

Thorn, C. F., Klein, T. E. and Altman, R. B. (2013) 'PharmGKB: the Pharmacogenomics Knowledge Base', *Methods in molecular biology (Clifton, N.J.)*, 1015, pp. 311–320. doi: 10.1007/978-1-62703-435-7_20.

Tiroch, K. A. *et al.* (2010) 'Protective effect of the CYP2C19 *17 polymorphism with increased activation of clopidogrel on cardiovascular events', *American Heart Journal*, 160(3), pp. 506–512. doi: https://doi.org/10.1016/j.ahj.2010.06.039.

Tishkoff, S. A. *et al.* (2009) 'The Genetic Structure and History of Africans and African

Americans', *Science*, 324(5930), pp. 1035 LP – 1044. doi: 10.1126/science.1172257.

Di Tommaso, P. *et al.* (2017) 'Nextflow enables reproducible computational workflows', *Nature Biotechnology*, 35(4), pp. 316–319. doi: 10.1038/nbt.3820.

Trontelj, J. *et al.* (2009) 'Effects of UGT1A1*28 polymorphism on raloxifene pharmacokinetics and pharmacodynamics', *British journal of clinical pharmacology*, 67(4), pp. 437–444. doi: 10.1111/j.1365-2125.2009.03363.x.

Turatti, L. *et al.* (2011) 'Short Communication: UGT1A1*28 Variant Allele Is a Predictor of Severe Hyperbilirubinemia in HIV-Infected Patients on HAART in Southern Brazil', *AIDS Research and Human Retroviruses*, 28(9), pp. 1015–1018. doi: 10.1089/aid.2011.0261.

Turner, R. M. *et al.* (2020) 'A Genome-wide Association Study of Circulating Levels of Atorvastatin and Its Major Metabolites', *Clinical Pharmacology & Therapeutics*, 108(2), pp. 287–297. doi: 10.1002/cpt.1820.

Ufer, M. (2005) 'Comparative Pharmacokinetics of Vitamin K Antagonists', *Clinical Pharmacokinetics*, 44(12), pp. 1227–1246. doi: 10.2165/00003088-200544120-00003.

Umamaheswaran, G., Kumar, D. K. and Adithan, C. (2014) 'Distribution of genetic polymorphisms of genes encoding drug metabolizing enzymes & drug transporters - a review with Indian perspective', *The Indian journal of medical research*, 139(1), pp. 27–65. Available at: https://pubmed.ncbi.nlm.nih.gov/24604039.

Uren, C. *et al.* (2016) 'Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries', *Genetics*, 204(1), pp. 303–314. doi: 10.1534/genetics.116.187369.

Vardhanabhuti, S. *et al.* (2015) 'Screening for UGT1A1 Genotype in Study A5257 Would Have Markedly Reduced Premature Discontinuation of Atazanavir for Hyperbilirubinemia', *Open forum infectious diseases*, 2(3), pp. ofv085–ofv085. doi: 10.1093/ofid/ofv085.

Vemula, P. K. *et al.* (2011) '4.421 - Self-Assembled Prodrugs', in Ducheyne, P. B. T.-C. B. (ed.). Oxford: Elsevier, pp. 339–355. doi: https://doi.org/10.1016/B978-0-08-055294-1.00135-5.

Venkataramanan, R. *et al.* (1995) 'Clinical Pharmacokinetics of Tacrolimus', *Clinical Pharmacokinetics*, 29(6), pp. 404–430. doi: 10.2165/00003088-199529060-00003.

Venter, W. D. F. *et al.* (2019) 'Dolutegravir plus Two Different Prodrugs of Tenofovir to Treat HIV', *New England Journal of Medicine*, 381(9), pp. 803–815. doi: 10.1056/NEJMoa1902824.

Ventola, C. L. (2011) 'Pharmacogenomics in clinical practice: reality and expectations', *P & T : a peer-reviewed journal for formulary management*, 36(7), pp. 412–450. Available at: https://pubmed.ncbi.nlm.nih.gov/21931473.

Veyrieras, J.-B. *et al.* (2008) 'High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation', *PLOS Genetics*, 4(10), p. e1000214. Available at: https://doi.org/10.1371/journal.pgen.1000214.

Visscher, P. M. *et al.* (2017) '10 Years of GWAS Discovery: Biology, Function, and Translation', *American journal of human genetics*, 101(1), pp. 5–22. doi: 10.1016/j.ajhg.2017.06.005.

Vogel, C. L. *et al.* (2002) 'Efficacy and Safety of Trastuzumab as a Single Agent in First-Line Treatment of HER2-Overexpressing Metastatic Breast Cancer', *Journal of Clinical Oncology*, 20(3), pp. 719–726. doi: 10.1200/JCO.2002.20.3.719.

Voight, B. F. *et al.* (2010) 'Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis', *Nature genetics*, 42(7), pp. 579–589. doi: 10.1038/ng.609.

Vourvahis, M. *et al.* (2019) 'No Clinical Impact of CYP3A5 Gene Polymorphisms on the Pharmacokinetics and/or Efficacy of Maraviroc in Healthy Volunteers and HIV-1-Infected Subjects', *Journal of clinical pharmacology*. 2018/09/07, 59(1), pp. 139–152. doi: 10.1002/jcph.1306.

Walmsley, S. L. *et al.* (2013) 'Dolutegravir plus Abacavir–Lamivudine for the Treatment of HIV-1 Infection', *New England Journal of Medicine*, 369(19), pp. 1807–1818. doi: 10.1056/NEJMoa1215541.

Wang, L., McLeod, H. L. and Weinshilboum, R. M. (2011) 'Genomics and drug response', *New England Journal of Medicine*, 364(12), pp. 1144–1153. doi: 10.1056/NEJMra1010600.

Wang, X. *et al.* (2014) 'Mapping of hepatic expression quantitative trait loci (eQTLs) in a Han Chinese population', *Journal of Medical Genetics*, 51(5), pp. 319 LP – 326. doi: 10.1136/jmedgenet-2013-102045.

Wei, G. *et al.* (2007) 'Acute liver failure in Sweden: etiology and outcome', *Journal of Internal Medicine*, 262(3), pp. 393–401. doi: 10.1111/j.1365-2796.2007.01818.x.

Wen, M.-S. *et al.* (2008) 'Prospective Study of Warfarin Dosage Requirements Based on CYP2C9 and VKORC1 Genotypes', *Clinical Pharmacology & Therapeutics*, 84(1), pp. 83–89. doi: 10.1038/sj.clpt.6100453.

Wienkers, L. C. and Heath, T. G. (2005) 'Predicting in vivo drug interactions from in vitro drug discovery data', *Nature Reviews Drug Discovery*, 4(10), pp. 825–833. doi: 10.1038/nrd1851.

Wilke, R. A. *et al.* (2007) 'Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges', *Nature reviews. Drug discovery*, 6(11), pp. 904–916. doi: 10.1038/nrd2423.

Wilke, R. A., Moore, J. H. and Burmester, J. K. (2005) 'Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage', *Pharmacogenetics and Genomics*, 15(6). Available at: https://journals.lww.com/jpharmacogenetics/Fulltext/2005/06000/Relative_impact_of_CYP3A_genotype_and_concomitant.7.aspx.

Willrich, M. A. V., Hirata, M. H. and Hirata, R. D. C. (2009) 'Statin regulation of CYP3A4 and CYP3A5 expression', *Pharmacogenomics*, 10(6), pp. 1017–1024. doi: 10.2217/pgs.09.42.

Willrich, M. A. V *et al.* (2008) 'CYP3A5∗3A allele is associated with reduced lowering-lipid response to atorvastatin in individuals with hypercholesterolemia', *Clinica Chimica Acta*, 398(1), pp. 15–20. doi: https://doi.org/10.1016/j.cca.2008.07.032.

Wilson, J. F. *et al.* (2001) 'Population genetic structure of variable drug response', (December). doi: 10.1038/ng761.

Wishart, D. S. *et al.* (2018) 'DrugBank 5.0: a major update to the DrugBank database for 2018', *Nucleic acids research*, 46(D1), pp. D1074–D1082. doi: 10.1093/nar/gkx1037.

Wolf, C. R., Smith, G. and Smith, R. L. (2000) 'Science, medicine, and the future: Pharmacogenetics', *BMJ (Clinical research ed.)*, 320(7240), pp. 987–990. doi: 10.1136/bmj.320.7240.987.

Wright, F. A. *et al.* (2014) 'Heritability and genomics of gene expression in peripheral blood', *Nature genetics*. 2014/04/13, 46(5), pp. 430–437. doi: 10.1038/ng.2951.

WRIGHT, S. (1949) 'THE GENETICAL STRUCTURE OF POPULATIONS', *Annals of Eugenics*, 15(1), pp. 323–354. doi: https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

Xie, H.-G. *et al.* (2004) 'Genetic variability in CYP3A5 and its possible consequences', *Pharmacogenomics*, 5(3), pp. 243–272. doi: 10.1517/phgs.5.3.243.29833.

Yagura, H. *et al.* (2017) 'Impact of UGT1A1 gene polymorphisms on plasma dolutegravir trough concentrations and neuropsychiatric adverse events in Japanese individuals infected with HIV-1', *BMC infectious diseases*, 17(1), p. 622. doi: 10.1186/s12879-017-2717-x.

Yamamoto, A. *et al.* (2015) 'Sulphation of acetaminophen by the human cytosolic sulfotransferases: a systematic analysis', *Journal of biochemistry*. 2015/06/11, 158(6), pp. 497–504. doi: 10.1093/jb/mvv062.

Yang, H.-C. *et al.* (2015) 'Glucose 6-phosphate dehydrogenase knockdown enhances IL-8 expression in HepG2 cells via oxidative stress and NF-κB signaling pathway', *Journal of inflammation (London, England)*, 12, p. 34. doi: 10.1186/s12950-015-0078-z.

Yasuda, S. U., Zhang, L. and Huang, S.-M. (2008) 'The Role of Ethnicity in Variability in Response to Drugs: Focus on Clinical Pharmacology Studies', *Clinical Pharmacology & Therapeutics*, 84(3), pp. 417–423. doi: 10.1038/clpt.2008.141.

You, D. *et al.* (2016) 'The Association between GSTM1, GSTT1 Genetic Variants and Gastric Carcinoma Susceptibility in Chinese: A Systematic Review Article', *Iranian journal of public health*, 45(9), pp. 1103–1117. Available at: https://pubmed.ncbi.nlm.nih.gov/27957455.

Youngster, I. *et al.* (2010) 'Medications and Glucose-6-Phosphate Dehydrogenase Deficiency', *Drug Safety*, 33(9), pp. 713–726. doi: 10.2165/11536520-000000000-00000.

Zanger, U. M. *et al.* (2008) 'Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation', *Analytical and Bioanalytical Chemistry*, 392(6), pp. 1093–1108. doi: 10.1007/s00216-008-2291-6.

Zanger, U. M., Raimundo, S. and Eichelbaum, M. (2004) 'Cytochrome P450 2D6: overview and update on pharmacology, genetics, biochemistry', *Naunyn-Schmiedeberg's Archives of*

Pharmacology, 369(1), pp. 23–37. doi: 10.1007/s00210-003-0832-2.

Zanger, U. M. and Schwab, M. (2013) 'Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation', *Pharmacology & Therapeutics*, 138(1), pp. 103–141. doi: https://doi.org/10.1016/j.pharmthera.2012.12.007.

Zhang, F. *et al.* (2009) 'Copy number variation in human health, disease, and evolution', *Annual review of genomics and human genetics*, 10, pp. 451–481. doi: 10.1146/annurev.genom.9.081307.164217.

Zhang, Hong *et al.* (2007) 'DPYD*5 gene mutation contributes to the reduced DPYD enzyme activity and chemotherapeutic toxicity of 5-FU', *Medical Oncology*, 24(2), pp. 251–258. doi: 10.1007/BF02698048.

Zhang, X. *et al.* (2013) 'Discordant associations between SLCO1B1 521T→C and plasma levels of ritonavir-boosted protease inhibitors in AIDS clinical trials group study A5146', *Therapeutic drug monitoring*, 35(2), pp. 209–216. doi: 10.1097/FTD.0b013e318280d0ad.

Zhao, W. *et al.* (2017) 'Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease', *Nature genetics*. 2017/09/04, 49(10), pp. 1450–1457. doi: 10.1038/ng.3943.

Zheng, Z. *et al.* (2020) 'QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes', *Nucleic Acids Research*, 48(D1), pp. D983–D991. doi: 10.1093/nar/gkz888.

Zhong, Y. *et al.* (2020) 'Discovery of novel hepatocyte eQTLs in African Americans', *PLoS Genetics*, 16(4), pp. 1–22. doi: 10.1371/journal.pgen.1008662.

Zhou, S.-F. *et al.* (2009) 'Insights into the substrate specificity, inhibitors, regulation, and polymorphisms and the clinical impact of human cytochrome P450 1A2', *The AAPS journal*. 2009/07/10, 11(3), pp. 481–494. doi: 10.1208/s12248-009-9127-y.

Zhou, S. F., Liu, J. P. and Chowbay, B. (2009) 'Polymorphism of human cytochrome P450 enzymes and its clinical impact', *Drug Metabolism Reviews*, 41(2), pp. 89–295. doi: 10.1080/03602530902843483.

Zhou, Y., Ingelman-Sundberg, M. and Lauschke, V. M. (2017) 'Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects', *Clinical pharmacology and therapeutics*. 2017/05/26, 102(4), pp. 688–700. doi: 10.1002/cpt.690.

Zhu, W.-Y. *et al.* (2016) 'Association of CYP2C19 Polymorphisms with the Clinical Efficacy of Clopidogrel Therapy in Patients Undergoing Carotid Artery Stenting in Asia', *Scientific Reports*, 6(1), p. 25478. doi: 10.1038/srep25478.

APPENDIX A. ETHICS CLEARANCE CERTIFICATE

R14/49 Ms JM Mathew

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M170756

NAME: (Principal Investigator)	Ms JM Mathew
DEPARTMENT:	School of Molecular and Cell Biology & Sydney Brenner Institute for Molecular Bioscience
PROJECT TITLE:	Identification of expression quantitative trail loci (eQTLs) and their functional impact on ADME genes
DATE CONSIDERED:	28/07/2017
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof M Ramsay & Dr D Mavri-Damelin
APPROVED BY:	Professor PE Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	16/10/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially reviewed in **July** and will therefore be due in the month of **July** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

09/12/2020
Date

APPENDIX B. PLAGIARISM DECLARATION

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Jenny Mary Mathew (Student number: 1689624) am a student registered for the degree of PhD in Bioinformatics in the academic year 2021.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature:  Date: 1-November-2021

APPENDIX C. TURNITIN REPORT

1689624:Thesis_content_only_v5.docx

ORIGINALITY REPORT

22%	14%	17%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.pharmgkb.org Internet Source	1%
2	www.ncbi.nlm.nih.gov Internet Source	<1%
3	"Omics for Personalized Medicine", Springer Science and Business Media LLC, 2013 Publication	<1%
4	www.nature.com Internet Source	<1%
5	royalsocietypublishing.org Internet Source	<1%
6	link.springer.com Internet Source	<1%
7	journals.plos.org Internet Source	<1%
8	www.medicines.org.uk Internet Source	<1%

APPENDIX D. PUBLISHED JOURNAL PAPER IN PHARMACOGENOMICS JOURNAL, FUTURE MEDICINE (SEPTEMBER 2021)

***UGT1A1* regulatory variant with potential effect on efficacy of HIV and cancer drugs commonly prescribed in South Africa**

Abstract:

Background: Despite the high disease burden of human immunodeficiency virus (HIV) infection and colorectal cancer (CRC) in South Africa (SA), treatment-relevant pharmacogenetic variants are understudied. **Materials and methods**: Using publicly available genotype and gene expression data, a bioinformatic pipeline was developed to identify liver expression quantitative trait loci (eQTLs). **Results**: A novel cis-eQTL, rs28967009, was identified for *UGT1A1,* which is predicted to upregulate *UGT1A1* expression thereby potentially affecting the metabolism of dolutegravir and irinotecan, which are extensively prescribed in SA for HIV and CRC treatment, respectively. **Conclusion**: As increased *UGT1A1* expression could affect the clinical outcome of dolutegravir and irinotecan treatment by increasing drug clearance, patients with the rs28967009A variant may require increased drug doses to reach therapeutic levels or should be prescribed alternative drugs.

Keywords: eQTL, *UGT1A1*, Dolutegravir, Irinotecan, South Africa

Introduction

Expression quantitative trait loci (eQTLs) are genetic variants that contribute to variation in gene expression levels in specific tissues or cells, thus contributing to the modulation of protein levels in cells [1]. eQTLs have been implicated in susceptibility to disease, and some are lead single nucleotide polymorphisms (SNPs) in genome-wide association studies for complex diseases. Understanding the effect of sequence variation on gene regulation can thus contribute to understanding human disease [2]. With relevance to this study, eQTLs also have the intrinsic potential to assist in the prediction of the clinical outcome of treatment should they affect the expression of genes that modulate drug absorption, distribution, metabolism and excretion (ADME genes). eQTLs of ADME genes can be correlated with the level of drug

efficacy and an individual's propensity towards developing adverse drug reactions (ADRs) [3]. Therefore, eQTLs play an important role in pharmacogenomics as it is practically more feasible to determine an individual's genotype as a proxy for gene expression, in an effort to ensure that a patient is prescribed drug therapy regimens best suited to their genetic makeup.

The frequencies of eQTLs of ADME genes vary both within and between populations. African ancestry populations show the most inter-population variation for known genetic variants and East Asian ancestry populations generally show the least [4]. There is evidence for significant population sub-structure within South Africa [5] that likely also applies to variants in ADME genes. As an example of ethnic differences, at the same dose, a two-fold higher systemic concentration of rosuvastatin, a drug administered for lowering low-density lipoprotein cholesterol (LDL) levels, is observed in individuals of East Asian ancestry compared to those of European ancestry. Hence, the recommended dose of rosuvastatin in East Asians is lower, at 5 mg, half the commonly prescribed dose for other populations [6]. Although this is likely caused by genetic variation, Lee at al. detected no associations with two *SLCO1B1* variants (A 388>G and T 521>C). A more recent study, however, showed that rosuvastatin levels were higher in individuals with the *SLCO1B1* 521C allele and the *ABCG2* 421A allele [7] suggesting a possible genetic aetiology. Data on genetic variation in ADME genes can therefore help clinicians and researchers to understand individual differences in sensitivity or resistance to certain drugs, thereby avoiding ADRs or ineffective treatment in patients and improving the quality of therapies [8].

Whilst genetic polymorphisms and environmental factors influence how effective a drug is and the individuals' propensity to develop ADRs, 90% of these genetic variants are in non-coding regions of the genome, which could have important regulatory functions [9]. Nevertheless, extensive in-depth knowledge of genetic variations, pre-existing clinical factors and environmental variables only explain a small proportion of variance in drug response and despite advances in understanding drug response variation, there is still a large gap in our knowledge of the causes of inter-individual differences. Changes in gene expression levels have been found to be a major mechanism that underlies drug response variation and susceptibility to disease. eQTLs of ADME genes, could therefore be used to test relevant pharmacogenetic associations with drugs that are metabolised by the corresponding

enzyme(s). Since eQTLs modulate gene expression levels, a pipeline that identifies eQTLs for ADME genes in relevant tissues could help to explain some of the variance in drug dose requirements that contribute to the currently unexplainable portion of variability in drug response and propensity to ADRs. The pipeline could be applied to data from any tissue type, but in this study, we focussed on the liver since ADME genes are abundantly expressed in this organ.

This study reports on the predicted effect of a novel cis-eQTL on drugs metabolized by uridine diphosphate glucuronosyl transferase1A1 encoded by *UGT1A1*, an ADME gene for a phase II drug metabolizing enzyme. UGT1A1 is involved in the metabolism and clearance of a range of drugs and as such has been the focus of recent pharmacogenomics studies since deviation from what is considered normal expression levels, can significantly impact drug response. We focused on dolutegravir and irinotecan as these drugs are commonly prescribed in South Africa (SA) for HIV and colorectal cancer (CRC), respectively. The 16 Southern African Development Community (SADC) countries have the largest number of people living with HIV/acquired immunodeficiency syndrome (AIDS) [10], and SA has the largest HIV epidemic in the world with 7.2 million HIV positive individuals [11]. CRC is the fourth most common and sixth most lethal cancer in SA [12,13]. Despite improved survival from antiretroviral therapy (ART) and chemotherapy, HIV and CRC remain serious causes of morbidity and mortality in SA. The potential impact of an African-specific variant of *UGT1A1* on HIV and CRC treatment is explored in this study.

Materials and Methods

Datasets used and Quality control (QC) done: The genotype and gene expression datasets, GSE39036 and GSE32504 respectively, used in this analysis were obtained from a study by Schröder *et al.*, 2013, and downloaded from NCBI's Gene Expression Omnibus (GEO) [14]. Prior to QC, this dataset had 149 samples, 318,237 genotyped single nucleotide polymorphisms (SNPs) and 15,439 gene expression transcripts. The genotype dataset from GSE39036 was imputed to increase the number of SNPs using the Michigan imputation server with the Phase 3 1000 Genomes Project reference panel. Post imputation, r^2 value (reflecting

imputation quality) greater than 0.8, minor allele count greater than one, SNPs with 2 alleles and SNPs with missing data of up to 60% were tolerated.

Ethics Approval: Ethics approval (M170756) was obtained from the University of the Witwatersrand, Johannesburg, South Africa. This study was conducted according to the principles of the Declaration of Helsinki.

Pipeline development: A Nextflow (https://www.nextflow.io/) bioinformatic pipeline was developed (https://github.com/phelelani/nf-eqtl) to identify tissue-specific eQTLs from GSE39036 and GSE32504. The pipeline takes as input four files (normalized gene expression file, genotype file, gene expression platform specific supplementary file with probe information and genotype platform specific supplementary file with SNP data) and generates PLINK input and binary files (namely PED, MAP, BED, BIM and FAM files). Sample and SNP QC is then performed on the genotype dataset using PLINK (http://zzz.bwh.harvard.edu/plink/). Samples with discordant sex information, missingness of more than 5%, extreme mean heterozygosity and population outliers based on principal component analysis (PCA), as well as duplicated or related samples, are removed during QC and excluded from analysis. For SNP QC, SNPs with more than 6.5% missing data, minor allele frequency (MAF) lower than 1% and Hardy-Weinberg equilibrium (HWE) p-value lower than 0.00001, are excluded from the analysis. The pipeline identifies sample outliers using a basic PCA plot based on the normalized gene expression dataset. Since samples in the genotype and gene expression datasets have different 'Sample IDs', the pipeline uses 'sample titles' to map and identify which samples must be removed thereby ensuring that only individual level matched genotype and gene expression data are included in the analysis. The pipeline was used to generate the five input files namely: gene expression file; gene location file; genotype file; SNP location file and covariate file, which are required to identify eQTLs from the datasets. Once created, the pipeline uses the R-package MatrixEQTL to generate the list of eQTLs (Shabalin, 2012). Cis-eQTLs are defined as being within 1 megabase (Mb) on either side of a gene's transcription start site (TSS) and trans-eQTLs are at least 5 Mb downstream or upstream of a gene's TSS or on a different chromosome. eQTLs were detected first using the post-QC genotype and gene expression data and then using imputed genotype data and gene

expression data. An identified cis-eQTL was only considered to be significant if it had a p-value <0.05 and a FDR value <0.05.

Bioinformatic analysis: The list of genes with associated eQTLs generated using our pipeline from both the datasets, GSE39036 and GSE32504, was searched for core ADME genes obtained from the PharmaADME consortium (http://pharmaadme.org/joomla/). An in-depth in-silico analysis was performed for all the significant cis-eQTLs identified. The first step was to find whether the chromosome coordinates obtained for each eQTL had rsIDs associated with them using the online tool SNPnexus. The GTEx database was then queried to identify whether the ADME gene associated with each cis-eQTL in question, is expressed in the liver and whether the eQTL has been previously identified. Allele frequencies of the eQTLs in different African populations were assessed using whole genome sequence (WGS) data from the 1000 Genomes Project (KGP) African populations and African Genome Variation Project (AGVP) WGS data (n=320). If present in African populations, LDlink was used to identify if the cis-eQTL is in linkage disequilibrium (LD) with variants in the gene whose expression it is associated with. Finally, drug label and variant annotations associated with the cis-eQTLs were derived from PharmGKB.

Results

During sample QC of the genotype dataset, no sex discordances, duplicates, related or samples with extreme mean heterozygosity were identified. Four samples had missingness of >5% (GSM954409, GSM954410, GSM954415, GSM954416) and three samples were identified as population outliers (GSM954398, GSM954391, GSM954346). These seven samples were excluded from further analysis. From a total of 318,237 SNPs, 20,652 were excluded from the analysis based on call rates, MAF and departure from HWE. QC of the gene expression dataset led to the removal of an additional six samples (GSM804707, GSM804739, GSM804673, GSM804662, GSM804715 and GSM804689) from further analysis as they were identified as outliers. Therefore, a total 136 samples were included for further analysis.

When the pipeline was run using GSE39036 and GSE32504, no cis-eQTLs were identified for the core ADME genes. However, when run using the imputed genotype dataset and GSE32504, six cis-eQTLs for core-ADME genes were identified (Table 1). The six core genes for which cis-eQTLs were identified, are all expressed in the liver, and none has previously been identified as eQTLs in GTEx, GeneCards and Ensembl.

From the six eQTLs identified, we found that only the cis-eQTL for *UGT1A1*, rs28967009, located 153,640 bps upstream from the TSS, is present in the African populations from the two WGS datasets (Table 2). The allele frequencies of the eQTL for *UGT1A1* in the different non-African populations from the KGP dataset are shown in Table 3. This cis-eQTL was present in 10 different African populations in varying allele frequencies (ranging from 0.025 to 0.125) and is generally lower in non-African populations. It is almost absent in European populations (except for the Finnish population at 0.030), and low in Asian populations (ranging from 0.005 to 0.059). The cis-eQTL rs28967009 was not in LD with the *UGT1A1* TA repeat allele variants (**28*, **37* and **36*) in African populations from the KGP (Yoruba in Nigeria (YRI), Luhya in Kenya (LWK), Gambian in Western Gambia (GWD), Mende in Sierra Leone (MSL), Esan in Nigeria (ESN), Americans of African ancestry in south-west United States of Africa (ASW) and African Caribbeans in Barbados (ACB)).The prediction of this SNP as a cis-eQTL is significant with a FDR of 0.006, a p-value of 5.73E-07 and a beta value of 1328. A beta value of 1328 indicates that the variant allele (allele A), is predicted to increase the expression of *UGT1A1* by 1328 times on average. A comprehensive list of variant annotations for *UGT1A1* is available in PharmGKB (https://www.pharmgkb.org/gene/PA420/variantAnnotation).

Discussion

This study identified rs28967009 as a cis-eQTL for uridine diphosphate glucuronosyl transferase1A1 (*UGT1A1*), a core ADME gene located at 2q37 that encodes an enzyme which plays a major role in the conjugation and subsequent elimination of xenobiotics, carcinogens and pharmaceutical drugs [15,16]. This cis-eQTL may have important consequences for the metabolism of dolutegravir and irinotecan, which are extensively prescribed in South Africa for HIV and colorectal cancer, respectively. Since the variant allele A, of the cis-eQTL is

associated with increased *UGT1A1* expression, it may affect the clinical outcome of dolutegravir and irinotecan treatment. Figure 1 illustrates the potential effect the variant allele A could have on *UGT1A1* expression and phenotypic outcome, in the presence of additional alleles that may affect the functional activity of the enzyme.

Irinotecan is widely prescribed in combination with other drugs to treat advanced/metastatic and/or recurrent colorectal cancer [17]. It is a prodrug that is metabolised by carboxylesterases in the liver to the active form SN-38, which is 100-1000 times more cytotoxic than irinotecan [18]. UGT1A1 is involved in the glucuronidation of SN-38 and results in the formation of SN-38 glucuronide, the excretable form of the drug [19]. Since rs28967009A, the *UGT1A1* variant allele of the cis-eQTL, correlates with greater *UGT1A1* expression levels, this variant is predicted to result in faster/increased glucuronidation of SN-38 to SN-38 glucuronide and this suggests that it is cleared from the body before SN-38 reaches therapeutic levels. Therefore, determining an optimal dose for irinotecan in individuals carrying this variant allele may be more challenging since it has a very narrow therapeutic range. Irinotecan treatment is generally limited by the high incidence of drug-toxicity due to high levels of SN-38 but would not be effective if the circulation levels were below the therapeutic range. The common adverse drug reactions (ADRs) of irinotecan include severe neutropenia, fever, asthenia and diarrhoea, caused by prolonged exposure to SN-38, and can be severe enough to lead to the prescription of a reduced drug dose or discontinuation of drug use [16,20]. The rate of severe ADRs associated with irinotecan treatment is around 20-25% [21] and approximately 7% of patients with severe neutropenia and fever following irinotecan treatment die from these complications [22,23]. In general, it is predicted that individuals with the cis-eQTL rs28967009A, will have a lower tendency to develop ADRs as the predicted increased expression of *UGT1A1* will result in the drug being cleared from the system faster, but it may also lead to treatment failure.

Dolutegravir is an anti-retroviral drug indicated for the treatment of HIV infection in combination with other ARTs. It is primarily metabolized in the liver by UGT1A1 to form dolutegravir glucuronide, which is the excretable form of the drug [24,25]. Dolutegravir is

widely used in SA and other developing countries in combination with tenofovir and lamivudine as first-line ART [26–28]. It is highly effective in reducing the HIV viral load thereby lowering the chance of HIV transmission and developing AIDS and HIV-related illnesses or complications like infections and cancer [29]. Since the cis-eQTL rs28967009A allele is predicted to increase *UGT1A1* expression, it may likely result in faster clearance of the drug from the body. Dose modification for dolutegravir could be recommended to avoid ineffective doses in individuals with the rs28967009A allele. Common ADRs associated with dolutegravir are neuropsychiatric adverse events (NPAEs) like mood changes, depression and anxiety and gastrointestinal adverse events like nausea and vomiting [30–33]. The presence of the cis-eQTL rs28967009A, is hypothesised to lower the chance of developing ADRs as the drug is likely to be cleared faster from the system because of the predicted increase in *UGT1A1* expression and/or may result in treatment failure because therapeutic levels of the drug are not reached.

As variants in the promoter and coding regions of *UGT1A1* are known to affect the expression level and activity of the enzyme, respectively, genetic testing of *UGT1A1* variants is recommended prior to administering any drugs metabolised by *UGT1A1*. Commonly known variants of *UGT1A1* are *UGT1A1*1*, **6*, **2*, **28*, *36, **37* and **60, with *1, *36*, **28 and *37* occurring in the gene promoter region. In enzyme assays, the *UGT1A1*1* and **60* variants are associated with normal UGT1A1 activity; *UGT1A1 *28* and **37* variants are associated with decreased *UGT1A1* expression; and *UGT1A1*36* with increased *UGT1A1* expression [34]. *UGT1A1*6* (glycine-71 to arginine) results in reduced enzyme activity and is predominantly found in people of East Asian ancestry [35] and is absent in European and African populations [36–39]. *UGT1A1*27* is a cytosine to adenine change in codon 229 (686C>A) that is in linkage disequilibrium with *UGT1A1*28* which results in reduced enzyme activity [40]. *UGT1A1*28* is commonly found in Africans with an allele frequency of 0.39-0.40 and in African-Americans with an allele frequency of 0.42–0.45 [41], with a predicted frequency of homozygotes of 0.16. Similarly, the frequency of *UGT1A1*28* homozygosity in Europeans is found to be in the range 0.09-0.16 [42]. The *UGT1A1*36* (increased *UGT1A1* expression) and **37* (decreased *UGT1A1* expression) alleles occur almost exclusively in populations of African origin, with estimated allele frequencies of 0.03–0.10 and 0.02–0.07, respectively [38,39,43]. The joint

presence of these variants on the same haplotype has not been investigated and the potential outcomes on enzyme activity would need to be investigated (Figure 1).

From a phenotypic viewpoint, depending on which two alleles of *UGT1A1* a person has, they are classified as extensive metabolizer (EM), intermediate metabolizer (IM) or poor metabolizer (PM). An extensive metabolizer has two functional *UGT1A1* alleles, two increased function alleles or a functional allele and an increased function allele. An intermediate metabolizer has a functional allele or an increased function allele with a decreased function allele. A poor metabolizer has two decreased function alleles [34]. Individuals with increased UGT1A1 enzyme activity have lower circulating drug concentrations and may only reach sub-therapeutic levels at standard dosage, requiring higher doses to achieve efficacy. UGT1A1 IMs and PMs are at higher risk of side effects as the drug remains in circulation for extended periods of time because of poor/reduced UGT1A1 enzyme activity. Drug dosage must be lowered in such individuals to prevent the occurrence of ADRs. With the aforementioned in mind, individuals with one or two copies of the rs28967009A allele identified in our study, are predicted to be extensive metabolisers (EMs), either requiring increased drug doses or alternative treatment due to an inability to achieve therapeutic levels. PharmGKB documents that the European Medicines Agency (EMA), Health Canada Santé Canada (HCSC), Swiss Agency of Therapeutic Products (Swissmedic) and United States Food and Drug Administration (US FDA) indicate that the starting dose of irinotecan must be reduced by at least one level in patients homozygous for the low function allele, *UGT1A1*28,* and in patients who have experienced hematologic toxicity with previous treatment to avoid/reduce risk for hematologic toxicity.

The effect of the variant allele A of the cis-eQTL rs28967009 on UGT1A1 PMs, IMs and EMs is difficult to predict as the presence of multiple and possibly opposing effects of regulatory variants on the same haplotype is unknown (Figure 1). In UGT1A1 PMs and IMs with at least one rs28967009A allele, it is expected that adverse drug reactions would be low as the predicted increase in UGT1A1 expression would clear the toxic active metabolites of drugs like irinotecan and dolutegravir faster, however this may reduce the therapeutic effect. In

UGT1A1 EMs with at least one rs28967009A allele, dosage levels of irinotecan and dolutegravir may need to be increased to reach therapeutic levels as a result of the rapid clearance of the drugs likely making them ineffective. If increased doses are still likely to lead to a failure of drug efficacy, an alternate treatment strategy to irinotecan and dolutegravir should be recommended.

In addition to dolutegravir and irinotecan, which currently have drug label annotations for *UGT1A1* variants, but not the one identified by our study, the efficacy of various other drugs is also affected by variants in *UGT1A1.* These include atazanavir and ritonavir that are used as combination ART for the treatment of HIV in South Africa [44]. Atazanavir inhibits UGT1A1, thereby preventing the glucuronidation and elimination of bilirubin. This causes hyperbilirubinemia and eventually jaundice and can culminate in discontinuation of atazanavir treatment. Risk for bilirubin-related discontinuation of atazanavir is highest among *UGT1A1* PMs [20,34,45]. Variants in *UGT1A1* are associated with increased risk of nephrolithiasis (rs8330G, rs1042640G, rs10929303T), increased likelihood of developing hyperbilirubinemia (rs887829T), increased chance of discontinuation of atazanavir treatment (rs887829T) and increased metabolism of atorvastatin (rs887829T), a drug used in the treatment of cardiovascular disease [46–49]. rs4148323A is associated with decreased metabolism of SN-38 in people with neoplasms, decreased overall survival (OS) and progression free survival (PFS) [50–52]. *UGT1A1*28* is associated with slower elimination of raloxifene (prescribed to reduce the risk of invasive breast cancer) and development of hyperbilirubinemia when treated with tranilast, an antiallergic medication [38,53,54]. *UGT1A1*6* allele is associated with lower clearance of etoposide (an anticancer agent) in people of African ancestry [55]. Since the novel cis-eQTL described in this paper, rs28967009A, is predicted to increase *UGT1A1* expression, its presence along with existing variants of *UGT1A1* that increase susceptibility to adverse drug reactions, may lower the risk of developing these adverse drug reactions. For example, as rs887829T is associated with increased metabolism of atorvastatin, an alternate drug may need to be prescribed to people who have both rs887829T and the A allele of the cis-eQTL as the drug is postulated to be cleared much faster due to increased *UGT1A1* expression.

A limitation of this study is that it is based on one genotype-gene expression dataset pair because of unavailability of other datasets in the public domain. Therefore, the findings of this analysis must be replicated by additional studies using more datasets. In addition, the effect of the variant allele A of the cis-eQTL rs28967009 needs to be clinically and functionally validated before being included in drug label annotations. If clinically validated, changes in drug dose or the use of alternate drugs could be recommended for this variant.

Conclusion

UGT1A1 is a core ADME gene that plays a major role in the metabolism of drugs like dolutegravir and irinotecan that are prescribed extensively in South Africa to treat HIV and CRC respectively. Therapeutic levels of the drugs irinotecan and dolutegravir may not be reached in extensive metabolizers with the rs28967009A variant in *UGT1A1*, which is present at appreciable frequencies in different African populations, including those in South Africa. Furthermore, UGT1A1 poor and intermediate metabolizers who have at least one rs28967009A allele may have a reduced likelihood of developing adverse drug reactions but may not reach therapeutic drug levels of the active compound of the drugs. Since dolutegravir is an ART that is used extensively in SA because of the high disease burden of HIV and because CRC is the fourth most common cancer in SA, understanding the effect of the novel cis-eQTL on the metabolism of drugs commonly used for these disorders is of high importance to attain optimal drug efficacy levels. Pharmacogenomic testing of this variant prior to administering dolutegravir or irinotecan could present a novel strategy to identify individuals who may be poor responders due to elevated drug clearance. As such, this study identifies a sub-population of Africans that may benefit from alternative therapies.

Acknowledgements

JM was supported by a grant holder linked bursary from the South African National Research Foundation. MR is a South African Research Chair in Genomics and Bioinformatics of African populations hosted by the University of the Witwatersrand, funded by the Department of Science and Technology, and administered by the National Research Foundation.

Figures

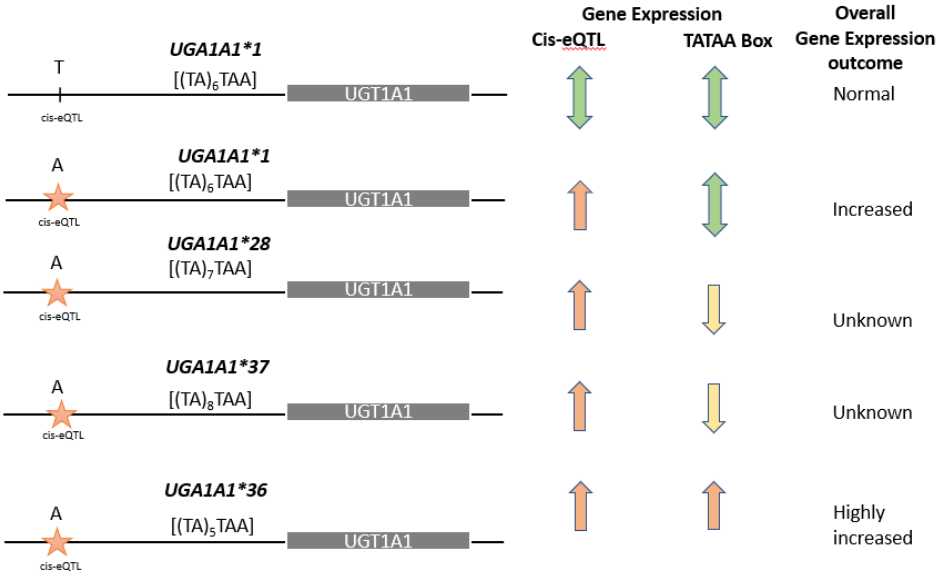


Figure 1: The effect of the variant allele A of the cis-eQTL, rs28967009, on *UGT1A1* gene expression in the presence of different star alleles on the same chromosome. The star allele nomenclature is based on biochemical assays of enzyme activity. The star alleles shown here, *UGT1A1*1, *28, *37* and **36* are well studied and occur in the regulatory TATAA Box, affecting the binding of transcription factors and the rate of transcription. These are haplotypes and the combinations of haplotypes in an individual will determine their metaboliser status.

[*UGT1A1*1* $[(TA)_6TAA]$ -->normal enzyme activity; *UGT1A1*36* $[(TA)_5TAA]$ -->increased activity (more enzyme produced); *UGT1A1*28* $[(TA)_7TAA]$ -->reduced activity (less enzyme produced); *UGT1A1*37* $[(TA)_8TAA]$ -->reduced activity]

Tables

Table 1: Cis-eQTLs for core ADME genes using the imputed genotype dataset and gene expression dataset

eQTL	rsID	Gene	β value	t-stat	p-value	FDR
10:97236373	rs933765367	*CYP2C19*	3238	5.74	5.64E-08	0.0009
7:100085809	rs143210517	*CYP3A5*	16158	5.82	3.76E-08	0.0006
1:97464825	rs114045355	*DPYD*	325	5.16	8.27E-07	0.0075
16:27948796	rs1000598999	*SULT1A1*	47	5.26	5.33E-07	0.0053
2:234515276	rs28967009	*UGT1A1*	1328	5.24	5.73E-07	0.0056
4:68462245	rs917897719	*UGT2B17*	3944	5.54	1.46E-07	0.0019

β value- Beta value; t-stat- t-statistic; p-value- calculated probability; FDR- false discovery rate

Table 2: Allele frequencies of the eQTL for UGT1A1 in different African populations as reported in the 1000 Genomes Project (KGP) and the African Human Genome Variation Project (AGVP)

eQTL for UGT1A1: rs28967009 (2:234515276)			
Dataset	Population	MAF	Sample size
KGP	GWD	0.031	113
	MSL	0.018	85
	ESN	0.061	99
	YRI	0.060	108
	LWK	0.081	99
	ASW	0.033	61
	ACB	0.042	96
AGVP	ZUL	0.080	100
	BAG	0.125	100
	ETH	0.025	120

GWD: Gambian in Western Divisions in the Gambia; MSL: Mende in Sierra Leone; ESN: Esan in Nigeria; YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; ASW: American's of African Ancestry in SW USA (admixed); ACB: African Caribbean in Barbados (admixed); ZUL: Zulu; BAG: Baganda; ETH: Ethiopia. ASW and ACB are admixed African populations.

Table 3: Allele frequencies of the eQTL for *UGT1A1* in non-African populations

eQTL for UGT1A1: rs28967009 (2:234515276)			
Dataset	Population	Sub-population	MAF
KGP	SAS	BEB	0.029
		ITU	0.039
		STU	0.044
		PJL	0.068
		GIH	0.029
	EUR	CEU	0
		IBS	0.009
		FIN	0.030
		TSI	0.009
		GBR	0
	AMR	CLM	0.037
		PEL	0.176
		MXL	0.062
		PUR	0.019
	EAS	CHB	0.010
		JPT	0.005
		CHS	0.038
		CDX	0.059
		KHV	0.035

SAS: South Asian; BEB: Bengali in Bangladesh; ITU: Indian Telugu in the UK; STU: Sri Lankan Tamil in the UK; PJL: Punjabi in Lahore, Pakistan; GIH: Gujarati Indian in Houston, Texas; EUR: European; CEU: Utah residents with Northern and Western European ancestry; IBS: Iberian populations in Spain; FIN: Finnish in Finland; TSI: Toscani in Italy ; GBR: British in England and Scotland ; AMR: American; CLM: Colombian in Medellin, Colombia ; PEL: Peruvian in Lima, Peru; MXL: Mexican Ancestry in Los Angeles, California; PUR: Puerto Rican in Puerto Rico; EAS: East Asian; CHB: Han Chinese in Beijing, China; JPT: Japanese

in Tokyo, Japan ; CHS: Southern Han Chinese, China; CDX: Chinese Dai in Xishuangbanna, China; KHV: Kinh in Ho Chi Minh City, Vietnam.

References

1. Nica AC, Dermitzakis ET. Expression quantitative trait loci: Present and future. *Philos. Trans. R. Soc. B Biol. Sci.* 368(1620) (2013).

2. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 28(10), 1353–1358 (2012).

3. Hovelson DH, Xue Z, Zawistowski M, *et al.* Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenet. Genomics*. 27(3), 89–100 (2017).

4. Ramos E, Doumatey A, Elkahloun AG, *et al.* Pharmacogenomics, ancestry and clinical decision making for global populations. *Pharmacogenomics J.* [Internet]. 14(3), 217–222 (2014). Available from: http://dx.doi.org/10.1038/tpj.2013.24.

5. Sengupta D, Choudhury A, Fortes-Lima C, *et al.* Genetic-substructure and complex demographic history of South African Bantu speakers. *bioRxiv* [Internet]. , 2020.08.11.243840 (2020). Available from: http://biorxiv.org/content/early/2020/08/11/2020.08.11.243840.abstract.

6. Lee H-K, Hu M, Lui SS, Ho C-S, Wong C-K, Tomlinson B. Effects of polymorphisms in ABCG2, SLCO1B1, SLC10A1 and CYP2C9/19 on plasma concentrations of rosuvastatin and lipid response in Chinese patients. *Pharmacogenomics* [Internet]. 14(11), 1283—1294 (2013). Available from: https://doi.org/10.2217/pgs.13.115.

7. Birmingham BK, Bujac SR, Elsby R, *et al.* Rosuvastatin pharmacokinetics and pharmacogenetics in Caucasian and Asian subjects residing in the United States. *Eur. J. Clin. Pharmacol.* [Internet]. 71(3), 329–340 (2015). Available from: https://doi.org/10.1007/s00228-014-1800-0.

8. Iskakova AN, Romanova AA, Aitkulova AM, Sikhayeva NS, Zholdybayeva E V., Ramanculov EM. Polymorphisms in genes involved in the absorption, distribution, metabolism, and excretion of drugs in the Kazakhs of Kazakhstan. *BMC Genet.* [Internet]. 17(1) (2016). Available from: http://dx.doi.org/10.1186/s12863-016-0329-x.

9. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta - Mol. Basis Dis.* [Internet]. 1842(10), 1910–1922 (2014). Available from: http://www.sciencedirect.com/science/article/pii/S0925443914000714.

10. Gona PN, Gona CM, Ballout S, *et al.* Burden and changes in HIV/AIDS morbidity and mortality in Southern Africa Development Community Countries, 1990–2017. *BMC Public Health* [Internet]. 20(1), 867 (2020). Available from: https://doi.org/10.1186/s12889-020-08988-9.

11. Satoh S, Boyer E. HIV in South Africa. *Lancet* [Internet]. 394(10197), 467 (2019). Available from: http://dx.doi.org/10.1016/S0140-6736(19)31634-4.

12. Brand M, Gaylard P, Ramos J. Colorectal cancer in South Africa: An assessment of disease presentation, treatment pathways and 5-year survival. *South African Med. J.* [Internet]. 108(2), 118 (2018). Available from: http://www.samj.org.za/index.php/samj/article/view/12188.

13. Herbst C, Miot JK, Moch SL, Ruff P. Colorectal Cancer (CRC) treatment and associated costs in the public sector compared to the private sector in Johannesburg, South Africa. *BMC Health Serv. Res.* [Internet]. 20(1), 290 (2020). Available from: https://doi.org/10.1186/s12913-020-05112-w.

14. Schröder A, Klein K, Winter S, *et al.* Genomics of ADME gene expression: Mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *Pharmacogenomics J.* 13(1), 12–20 (2013).

15. Schulz C, Boeck S, Heinemann V, Stemmler H-J. UGT1A1 genotyping: a predictor of irinotecan-associated side effects and drug efficacy? *Anticancer. Drugs* [Internet]. 20(10), 867—879 (2009). Available from: https://doi.org/10.1097/CAD.0b013e328330c7d2.

16. Barbarino JM, Haidar CE, Klein TE, Altman RB. PharmGKB summary: very important pharmacogene information for UGT1A1. *Pharmacogenet. Genomics* [Internet]. 24(3), 177–183 (2014). Available from: https://pubmed.ncbi.nlm.nih.gov/24492252.

17. Ramesh M, Ahlawat P, Srinivas NR. Irinotecan and its active metabolite, SN-38: review of bioanalytical methods and recent update from clinical pharmacology perspectives. *Biomed. Chromatogr.* [Internet]. 24(1), 104–123 (2010). Available from: https://doi.org/10.1002/bmc.1345.

18. Mathijssen RHJ, van Alphen RJ, Verweij J, *et al.* Clinical Pharmacokinetics and Metabolism of Irinotecan (CPT-11). *Clin. Cancer Res.* [Internet]. 7(8), 2182 LP – 2194 (2001). Available from: http://clincancerres.aacrjournals.org/content/7/8/2182.abstract.

19. Dias MM, McKinnon RA, Sorich MJ. Impact of the UGT1A1*28 allele on response to irinotecan: a systematic review and meta-analysis. *Pharmacogenomics* [Internet]. 13(8), 889–899 (2012). Available from: https://doi.org/10.2217/pgs.12.68.

20. Lankisch TO, Schulz C, Zwingers T, *et al.* Gilbert's Syndrome and Irinotecan Toxicity: Combination with UDP-Glucuronosyltransferase 1A7 Variants Increases Risk. *Cancer Epidemiol. Biomarkers & Prev.* [Internet]. 17(3), 695 LP – 701 (2008). Available from: http://cebp.aacrjournals.org/content/17/3/695.abstract.

21. Tam VC, Rask S, Koru-Sengul T, Dhesy-Thind S. Generalizability of toxicity data from oncology clinical trials to clinical practice: toxicity of irinotecan-based regimens in patients with metastatic colorectal cancer. *Curr. Oncol.* [Internet]. 16(6), 13–20 (2009). Available from: https://pubmed.ncbi.nlm.nih.gov/20016742.

22. Obradovic M, Mrhar A, Kos M. Cost–effectiveness of UGT1A1 genotyping in second-line, high-dose, once every 3 weeks irinotecan monotherapy treatment of colorectal cancer. *Pharmacogenomics* [Internet]. 9(5), 539–549 (2008). Available from: https://doi.org/10.2217/14622416.9.5.539.

23. Liu X, Cheng D, Kuang Q, Liu G, Xu W. Association of UGT1A1*28 polymorphisms with irinotecan-induced toxicities in colorectal cancer: a meta-analysis in Caucasians. *Pharmacogenomics J.* [Internet]. 14(2), 120–129 (2014). Available from: https://pubmed.ncbi.nlm.nih.gov/23529007.

24. Castellino S, Moss L, Wagner D, *et al.* Metabolism, excretion, and mass balance of the HIV-1 integrase inhibitor dolutegravir in humans. *Antimicrob. Agents Chemother.* [Internet]. 57(8), 3536–3546 (2013). Available from: https://pubmed.ncbi.nlm.nih.gov/23669385.

25. Yagura H, Watanabe D, Kushida H, *et al.* Impact of UGT1A1 gene polymorphisms on plasma dolutegravir trough concentrations and neuropsychiatric adverse events in Japanese individuals infected with HIV-1. *BMC Infect. Dis.* [Internet]. 17(1), 622 (2017). Available from: https://pubmed.ncbi.nlm.nih.gov/28915895.

26. Walmsley SL, Antela A, Clumeck N, *et al.* Dolutegravir plus Abacavir–Lamivudine for the Treatment of HIV-1 Infection. *N. Engl. J. Med.* [Internet]. 369(19), 1807–1818 (2013). Available from: https://doi.org/10.1056/NEJMoa1215541.

27. Moorhouse MA, Carmona S, Davies N, *et al.* Southern African HIV Clinicians Society Guidance on the use of dolutegravir in first-line antiretroviral therapy. *South. African J. HIV Med. Vol 19, No 1 (2018)DO - 10.4102/sajhivmed.v19i1.917* [Internet]. (2018). Available from: https://sajhivmed.org.za/index.php/hivmed/article/view/917/1276.

28. Venter WDF, Moorhouse M, Sokhela S, *et al.* Dolutegravir plus Two Different Prodrugs of Tenofovir to Treat HIV. *N. Engl. J. Med.* [Internet]. 381(9), 803–815 (2019). Available from: https://doi.org/10.1056/NEJMoa1902824.

29. Spano J-P, Poizot-Martin I, Costagliola D, *et al.* Non-AIDS-related malignancies: expert consensus review and practical applications from the multidisciplinary CANCERVIH Working Group. *Ann. Oncol.* [Internet]. 27(3), 397–408 (2016). Available from: https://doi.org/10.1093/annonc/mdv606.

30. Elzi L, Erb S, Furrer H, *et al.* Adverse events of raltegravir and dolutegravir. *AIDS* [Internet]. 31(13), 1853–1858 (2017). Available from: https://pubmed.ncbi.nlm.nih.gov/28692533.

31. Lepik KJ, Yip B, Ulloa AC, *et al.* Adverse drug reactions to integrase strand transfer inhibitors. *AIDS* [Internet]. 32(7) (2018). Available from: https://journals.lww.com/aidsonline/Fulltext/2018/04240/Adverse_drug_reactions_to_integrase_strand.9.aspx.

32. Cuzin L, Pugliese P, Katlama C, *et al.* Integrase strand transfer inhibitors and neuropsychiatric adverse events in a large prospective cohort. *J. Antimicrob. Chemother.* [Internet]. 74(3), 754–760 (2019). Available from: https://doi.org/10.1093/jac/dky497.

33. Llibre JM, Montoliu A, Miró JM, *et al.* Discontinuation of dolutegravir, elvitegravir/cobicistat and raltegravir because of toxicity in a prospective cohort. *HIV Med.* [Internet]. 20(3), 237–247 (2019). Available from: https://doi.org/10.1111/hiv.12710.

34. Gammal RS, Court MH, Haidar CE, *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for UGT1A1 and Atazanavir Prescribing. *Clin. Pharmacol. Ther.* [Internet]. 99(4), 363–369 (2016). Available from: https://pubmed.ncbi.nlm.nih.gov/26417955.

35. Teh LK, Hashim H, Zakaria ZA, Salleh MZ. Polymorphisms of UGT1A1*6, UGT1A1*27 & UGT1A1*28 in three major ethnic groups from Malaysia. *Indian J. Med. Res.* [Internet]. 136(2), 249–259 (2012). Available from: https://pubmed.ncbi.nlm.nih.gov/22960892.

36. Dai X, Wu C, He Y, *et al.* A Genome-Wide Association Study for Serum Bilirubin Levels and Gene-Environment Interaction in a Chinese Population. *Genet. Epidemiol.* [Internet]. 37(3), 293–300 (2013). Available from: https://doi.org/10.1002/gepi.21711.

37. Liu JY, Qu K, Sferruzza AD, Bender RA. Distribution of the UGT1A1*28 polymorphism in Caucasian and Asian populations in the US: a genomic analysis of 138 healthy individuals. *Anticancer. Drugs* [Internet]. 18(6) (2007). Available from: https://journals.lww.com/anti-cancerdrugs/Fulltext/2007/07000/Distribution_of_the_UGT1A1_28_polymorphism_in.9.aspx.

38. Palomaki GE, Bradley LA, Douglas MP, Kolor K, Dotson WD. Can UGT1A1 genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with irinotecan? An evidence-based review. *Genet. Med.* [Internet]. 11(1), 21–34 (2009). Available from: https://pubmed.ncbi.nlm.nih.gov/19125129.

39. Dean L. Irinotecan Therapy and UGT1A1 Genotype Drug : Irinotecan Irinotecan is used to treat colorectal cancer , which is the third most common cancer worldwide (6). It is often. (MD), 1–12 (2018).

40. Fukuda M, Suetsugu T, Shimada M, *et al.* Prospective study of the UGT1A1*27 gene polymorphism during irinotecan therapy in patients with lung cancer: Results of Lung Oncology Group in Kyusyu (LOGIK1004B). *Thorac. Cancer* [Internet]. 7(4), 467–472 (2016). Available from: https://doi.org/10.1111/1759-7714.12360.

41. Sukasem C, Atasilp C, Chansriwong P, Chamnanphon M, Puangpetch A, Sirachainan E. Development of Pyrosequencing Method for Detection of UGT1A1 Polymorphisms in Thai Colorectal Cancers. *J. Clin. Lab. Anal.* 30(1), 84–89 (2016).

42. Marques SC, Ikediobi ON. The clinical application of UGT1A1 pharmacogenetic testing: gene-environment interactions. *Hum. Genomics* [Internet]. 4(4), 238–249 (2010). Available from: https://pubmed.ncbi.nlm.nih.gov/20511137.

43. Beutler E, Gelbart T, Demina A. Racial variability in the UDP-glucuronosyltransferase 1 (UGT1A1) promoter: a balanced polymorphism for regulation of bilirubin metabolism? *Proc. Natl. Acad. Sci. U. S. A.* [Internet]. 95(14), 8170–8174 (1998). Available from: https://pubmed.ncbi.nlm.nih.gov/9653159.

44. Achenbach CJ, Darin KM, Murphy RL, Katlama C. Atazanavir/ritonavir-based combination antiretroviral therapy for treatment of HIV-1 infection in adults. *Future Virol.* [Internet]. 6(2), 157–177 (2011). Available from: https://pubmed.ncbi.nlm.nih.gov/21731578.

45. Turatti L, Sprinz E, Lazzaretti RK, *et al.* Short Communication: UGT1A1*28 Variant Allele Is a Predictor of Severe Hyperbilirubinemia in HIV-Infected Patients on HAART in Southern Brazil. *AIDS Res. Hum. Retroviruses* [Internet]. 28(9), 1015–1018 (2011). Available from: https://doi.org/10.1089/aid.2011.0261.

46. Johnson DH, Venuto C, Ritchie MD, *et al.* Genomewide association study of atazanavir pharmacokinetics and hyperbilirubinemia in AIDS Clinical Trials Group protocol A5202. *Pharmacogenet. Genomics* [Internet]. 24(4), 195–203 (2014). Available from: https://pubmed.ncbi.nlm.nih.gov/24557078.

47. Nishijima T, Tsuchiya K, Tanaka N, *et al.* Single-nucleotide polymorphisms in the UDP-glucuronosyltransferase 1A-3' untranslated region are associated with atazanavir-induced nephrolithiasis in patients with HIV-1 infection: a pharmacogenetic study. *J. Antimicrob. Chemother.* [Internet]. 69(12), 3320–3328 (2014). Available from: https://doi.org/10.1093/jac/dku304.

48. Vardhanabhuti S, Ribaudo HJ, Landovitz RJ, *et al.* Screening for UGT1A1 Genotype in Study A5257 Would Have Markedly Reduced Premature Discontinuation of Atazanavir for Hyperbilirubinemia. *Open forum Infect. Dis.* [Internet]. 2(3), ofv085–ofv085 (2015). Available from: https://pubmed.ncbi.nlm.nih.gov/26180834.

49. Turner RM, Fontana V, Zhang JE, *et al.* A Genome-wide Association Study of Circulating Levels of Atorvastatin and Its Major Metabolites. *Clin. Pharmacol. Ther.* [Internet]. 108(2), 287–297 (2020). Available from: https://doi.org/10.1002/cpt.1820.

50. Han J-Y, Lim H-S, Shin ES, *et al.* Comprehensive Analysis of UGT1A Polymorphisms Predictive for Pharmacokinetics and Treatment Outcome in Patients With Non–Small-Cell Lung Cancer Treated With Irinotecan and Cisplatin. *J. Clin. Oncol.* [Internet]. 24(15), 2237–2244 (2006). Available from: https://doi.org/10.1200/JCO.2005.03.0239.

51. Jada SR, Lim R, Wong CI, *et al.* Role of UGT1A1*6, UGT1A1*28 and ABCG2 c.421C>A polymorphisms in irinotecan-induced neutropenia in Asian cancer patients. *Cancer Sci.* [Internet]. 98(9), 1461–1467 (2007). Available from: https://doi.org/10.1111/j.1349-7006.2007.00541.x.

52. Takano M, Kato M, Yoshikawa T, *et al.* Clinical Significance of UDP-Glucuronosyltransferase 1A1*6 for Toxicities of Combination Chemotherapy with Irinotecan and Cisplatin in Gynecologic Cancers. *Oncology* [Internet]. 76(5), 315–321 (2009). Available from: https://www.karger.com/DOI/10.1159/000209335.

53. Danoff TM, Campbell DA, McCarthy LC, *et al.* A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia. *Pharmacogenomics J.* [Internet]. 4(1), 49–53 (2004). Available from: https://doi.org/10.1038/sj.tpj.6500221.

54. Trontelj J, Marc J, Zavratnik A, Bogataj M, Mrhar A. Effects of UGT1A1*28 polymorphism on raloxifene pharmacokinetics and pharmacodynamics. *Br. J. Clin. Pharmacol.* [Internet]. 67(4), 437–444 (2009). Available from: https://pubmed.ncbi.nlm.nih.gov/19371317.

55. Kishi S, Yang W, Boureau B, *et al.* Effects of prednisone and genetic polymorphisms on etoposide disposition in children with acute lymphoblastic leukemia. *Blood* [Internet]. 103(1), 67–72 (2004). Available from: https://doi.org/10.1182/blood-2003-06-2105.