UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

ADDRESSING CANCER TREATMENT IN AN AFRICAN SETTING: A BIOINFORMATICS ANALYSIS OF PHARMACOGENOMICALLY RELEVANT VARIANTS

DISSERTATION FOR A MASTERS DEGREE BY RESEARCH:

A Dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of

Master of Science in Medicine

by

Jorge da Rocha

Student Number: 670830

Supervisor: Prof Michèle Ramsay

Co-supervisor: Dr Zané Lombard

Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Johannesburg, June 2018

Declaration

I, Jorge da Rocha, declare that this Dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

(Signature of candidate)

On this the 6th day of June 2018 in Johannesburg

Acknowledgements

To my supervisors, Michèle and Zané, thank you for your unwavering support and guidance. The opportunities and lessons you have provided have helped me grow more than I ever thought I could. Thank you for helping on this path and I look forward to more research endeavors with you in future. To the examiners, thank you for your valuable input and review. These helped raise the standard of this work.

To my mom, you have raised our family from the ground up to exceptional heights. You are our provider, our rock, and our inspiration. I know you must be proud of us but know that we are so proud of you too.

Ao meu Vovô. Que sempre teve grande esperanças de mim. Obrigado pelo seu apoio e por me dar tudo que eu precisava para atinger os meus sonhos.

To my dearest sisters, Nicole and Claudia, every day I know that I am blessed to have you. You are always in my thoughts, no matter where we are in the world. Thank you for all your support and guidance. Alex and JP, your brotherly support is epic, and I know you will be there no matter what.

To Liesl, Shaun, and Phelelani, Jean-Tristan, and all the other wonderful colleagues at the SBIMB, you are like a second family to me now, and your help and support will always be remembered. Ananyo and Dhriti, you are like research parents, thank you for all you do for us at the SBIMB. Yusuf, Jocelyn, Mmatshepo and Stuart, our fearless project managers, you keep us all going.

To Mahtaab, Heather, Michaella and Andy, and all my other great friends, thank you for being there for me and for all your encouraging words.

To my darling Micke, you are the best partner anyone could wish for. Thank you for all you do and have done, from proofreading pages upon pages, to hearing my excited, lengthy and complex ramblings, you have been there with me all the way. I look forward to always achieving new heights together with you.

Financial Contributions which supported this work:

Thank you to:

The National Research Foundation of South Africa, for the grant-holder linked bursary.

The SBIMB and Prof Michèle, for the grants which assisted with travel and other expenses.

The Wellcome Trust Genome Center, for assistance with travel to attend the Next Generation Sequencing Bioinformatics Course, Hinxton, United Kingdom

H3ABionet, for the travel award to attend the Precision Medicine Conference and Workshop, Cape Town, South Africa.

University of the Witwatersrand, for the Postgraduate Merit Award to assist with fees.

List of Presentations

Precision Medicine Conference and Workshop hosted by H3ABionet, April 2016, University of Cape Town (UCT), Cape Town, South Africa

Poster presentation, Addressing Cancer Treatment in An African Setting: A Bioinformatics Analysis of Pharmacogenomically Relevant Variants

Wits Faculty of Health Sciences Research Day, September 2016, University of the Witwatersrand, Johannesburg, South Africa

Poster presentation, Addressing Cancer Treatment in An African Setting: A Bioinformatics Analysis of Pharmacogenomically Relevant Variants

Next Generation Sequencing Bioinformatics Course hosted by the Wellcome Genome Campus, October 2016, University of Cambridge, Hinxton, United Kingdom

Mini-Presentation, Addressing Cancer Treatment in An African Setting: A Bioinformatics Analysis of Pharmacogenomically Relevant Variants

Molecular Biosciences Research Tract (MBRT) Research Day, November 2016, University of the Witwatersrand, Johannesburg, South Africa

Poster presentation, Addressing Cancer Treatment in An African Setting: A Bioinformatics Analysis of Pharmacogenomically Relevant Variants

South African Society for Human Genetics (SASHG) 2017 Conference, August 2017, Durban, South Africa

Poster presentation, Addressing Cancer Treatment in An African Setting: A Bioinformatics Analysis of Pharmacogenomically Relevant Variants

Abstract

Cancer is a critical health burden in Africa, and mortality rates are rising rapidly. Treatments are severe and expensive, and often cause adverse-drug-reactions (ADRs). These may be attributed to variants in genes that encode proteins involved in key drug pathways. Most pharmacogenomics (PGx) research has been done in populations of European, Asian and African-American ancestry, with sparse data from African populations. Thirteen genes linked with ADRs to medicines used for treating major cancer types were identified: *ABCB1, DPYD, TYMS, CYP19A1, GSTP1, CYP1B1, CYP3A4, CYP3A5, ESR1, CYP2D6, SLC19A1* and *XRCC1/5*. Public domain whole-genome-sequencing data from the 1000 Genomes Project and the African Genome Variation Project were mined to assess eight African populations. Functional annotation was performed with a series of bioinformatics-based scoring tools to assess potential likelihood of deleterious impact. Variants of high likelihood of deleterious impact, including some novel variants were identified. These novel variants, however, are rare and require further validation. Two key African specific variants were identified: the *CYP3A5* frameshift variant, rs41303343, which is highly likely to knockout gene function, and the *CYP2D6* missense variant rs59421388, which was scored highly likely deleterious by all tools. Both variants are common in Africans, at frequencies above 10%, but lack clinical investigation into their PGx impact. For missense variants with known PGx effects, such as *CYP2D6* - rs1065852 and *DPYD* - rs2297595, African frequencies are significantly distinct from global populations ($p < 1x10^{-4}$) (with rs1065852 less common, and rs2297595 more common in Africans respectively). There is also variation within Africa, with the rs1065852 A allele being more common in West Africans, while the rs2297595 C allele is more common in East Africans. These data indicate that guidelines for cancer drug safety based on African data is essential for use in Africa, and novel region-specific guidelines should be developed to ensure that Africans could benefit for a personalized medicine approach.

Table of Contents

Nomenclature, List of Abbreviations and Symbols

ADR – adverse drug reaction

AGVP – African Genome Variation Project

BAM – binary alignment map

CPIC Clinical Pharmacogenomics Implementation Consortium

DNA – deoxyribose nucleic acid

FDA – United States Food and Drug Administration

KGP – The 1000 Genomes Project

LoF – loss of function

NCD – non-communicable disease

NGS – next generation sequencing

PDB – Protein Data Bank

PGx – pharmacogenomics

SNV/s – single nucleotide variant/s

UTR/s – untranslated Region/s

VCF – variant call file

WHO – World Health Organisation

WGS – whole genome sequencing

Gene list

ABCB1 – ATP-Binding Cassette sub-family B member 1

CYP19A1 – Human Aromatase

CYP1B1 – Cytochrome P450 Family 1 Subfamily B Member 1

CYP2D6 – Cytochrome P450 Family 2 Subfamily D Member 6

CYP3A4 – Cytochrome P450 Family 3 Subfamily A Member 4

CYP3A5 – Cytochrome P450 Family 3 Subfamily A Member 5

DPYD – Dihydropyrimidine Dehydrogenase

ESR1 – Oestrogen Receptor 1

GSTP1 – Glutathione S-Transferase Pi 1

SLC19A1 – Solute Carrier Family 19 Member 1, encodes RFC1 - Reduced Folate Carrier 1

TYMS – Thymidylate Synthase

XRCC1 – X-Ray Repair Cross Complementing 1

XRCC5 – X-Ray Repair Cross Complementing 5

Bioinformatics Toolsets:

CADD – Combined Annotation Dependent Depletion (CADD)

CAROL – Combined Annotation scoRing toOL

Condel – Consensus deleteriousness score of missense mutations

FATHMM-MKL – Functional Analysis through Hidden Markov Models – MKL

LOFTEE – Loss-Of-Function Transcript Effect Estimator

SIFT – Sorting Tolerant from Intolerant

VEP – Variant Effect Predictor

Population Names

ACB – African Caribbean, Barbados

ASW – Americans of African Ancestry, United States of America

BAG – Bagandan, Uganda

ESN – Esan, Nigeria

ETH – Ethiopian, Ethiopia

GWD – Gambian, The Gambia

LWK – Luhya, Kenya

MSL – Mende, Sierra Leone

YRI – Yoruba, Nigeria

ZULU – Zulu, South Africa

Super-Population Groups

AMR – Aggregate KGP Admixed Americans

Cont. AFR – Aggregate Continental Africans

EAS – Aggregate KGP East Asians

EUR – Aggregate KGP Europeans

SAS – Aggregate KGP South Asians

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 1: Introduction

Africa is the continent most severely affected by disease. The main contribution to disease burden on the continent is caused by communicable diseases. Although these diseases are devastating and prevalence is changing rapidly in a modern Africa, trends indicate that non-communicable diseases (NCDs) are set to increase (WHO, 2014), similar to what has occurred in developed countries (Arena *et al.*, 2015). These diseases are complex and are the result of multiple environmental and genetic factors. NCDs are often chronic, and require multiple visits to treatment centres and have high costs for tests and medications, which can place a large financial burden on communities already struggling with poverty (Mayosi *et al.*, 2009). When considering the rising rates of NCDs in Africa and the high financial burdens they place on patients, it is clear that there is a need for new, advanced strategies for prevention, diagnosis and treatment for these diseases across the continent.

The major NCDs affecting Africans are cardiovascular disease, diabetes and cancer (WHO, 2014). Cancer poses severe difficulties to African populations due to the lack of availability of medical treatment and high cost of drugs. Although varying by country, cancer is the second most common NCD related cause of death in Africa, the most common being cardiovascular disease (WHO, 2014). The World Health Organisation (WHO) predicts that Africa will bear an extreme increase in cancer related deaths over the next fifteen years, as shown in Table 1.1.

Table 1. 1 Cancer Mortality predictions for 2030 by the WHO for the African Region

	Mortality* 2015		Mortality* 2030	
Cancer Type	**Males**	**Females**	**Males**	**Females**
Mouth and oropharynx cancers	11556	5574	21877	9291
Oesophageal cancer	20607	9956	39748	15746
Stomach cancer	12991	10243	25589	18084
Colon and rectum cancers	16176	12226	30769	20558
Liver cancer	40827	18889	79697	32180
Pancreas cancer	4679	4391	9100	7480
Trachea, bronchus, lung cancers	16083	5977	33951	13250
Melanoma and other skin cancers	3600	3440	6683	5823
Breast cancer	0	50836	0	108299
Cervix uteri cancer	0	60661	0	100737
Corpus uteri cancer	0	2422	0	4099
Ovary cancer	0	10154	0	17276
Prostate cancer	35009	0	72161	0
Bladder cancer	7846	3069	15170	5147
Lymphomas, multiple myeloma	24527	16665	45541	28182
Leukemia	10164	7128	18338	11893
Other malignant neoplasms	50578	44754	94292	75920
Total:	**254645**	**266385**	**492915**	**473962**

**Predictions of Mortality performed on by the WHO based on the mortality data for 2013 (accessed at: http://www.who.int/healthinfo/global_burden_disease/GHE_DthWHOReg6_Proj_2015_2030.xls?ua=1 on the 20/04/2016)*

Based on the predictions (Table 1.1), cancer rates for most types are likely to almost double in the next 13 years. The cancer mortality profiles for males and females can be clearly distinguished by specific types. Breast and cervical cancers are the largest contributors to mortality in females. Liver, prostate, and oesophageal cancers are the largest contributors for males. It is interesting that cancers that occur in both sexes, such as liver, oesophageal, and lung cancers disproportionately affect males. Other cancer types such as lymphomas, melanomas and pancreatic cancer affect both sexes in similar profiles. These cancers were also noted to be causing a large number of deaths in individuals that are between 30 and 49 years old (WHO, 2014).

Breast cancers are predicted to overtake cervical cancers, causing the greatest mortality in women (Table 1.1). Breast cancer is highly prevalent in South Africa, and it has been reported that black South African women have a greater risk of aggressive subtypes (Dickens *et al.*, 2014). Liver- and prostate cancers will still be the most common in males. The WHO also predicts that the portion

of younger individuals afflicted will rise (WHO, 2014), greatly affecting burden, especially when considering that most cancer types require treatment over many years, often leaving the patient unable to work during treatment (de-Graft Aikins *et al.*, 2010). These losses of working years may be due to the fact that cancer treatments, such as chemotherapies and radiation treatment, are severe and produce a large number of side effects. These side effects or adverse drug reactions (ADRs) are variable between individuals, and are noted to vary between individuals from different populations or ethnicities (Patel, 2015). Relevant data pertaining to known adverse drug reactions related to ethnicity are available from the United States Food and Drug Administration (FDA). The FDA is a major regulator of the food and pharmaceutical industry to ensure safety and optimal treatments. Ethnicity-related drug warnings and biomarker labels are available to the public (FDA, 2016). Of the drugs that the FDA has labelled for potential ADRs, based on the ethnicity of the patient, the majority are for oncology-related treatments which highlights the need to address cancer drug safety, especially in African populations that have not yet been well characterised.

1.1 Pharmacogenomics

Pharmacogenomics (PGx) is an emerging research field that combines the fields of pharmacology with human genetics in order to better understand the relationship between the individual, their genes, and their response to specific drugs (Pirmohamed, 2001). The outcomes of this field will contribute to precision medicine, i.e. medical care that has high therapeutic efficacy while having minimalized side-effects for particular groups of patients, who are identified by use of genetic profiling. In order to achieve this outcome, the relationship between the drug and the individual's genotype is investigated. The genotypes involved are often found within and near to genes which encode drug metaboliser enzymes, drug transporters or the therapeutic target of the drug. Because of the unique relationship between a specific drug and the gene involved, they are often grouped as "drug-gene pairs". The relationship between drug and body are measured in terms of pharmacokinetics (i.e. what the body does to the drug), in which genes play large roles, and in pharmacodynamics (i.e. what the drug does to the body), which concerns the drug's mechanism of action (Meibohm and Derendorf, 1997). Drug-gene pairs are documented for various diseases and disease types, and are stored in highly curated databases, such as the PharmGKB database (Whirl-Carrillo *et al.*, 2012). This database is particularly well curated, and it has multiple layers of data available for researchers to mine and compare. These layers range from genetic information

related to drug types, to sets of drugs and genes involved and finally clinical annotation and guidelines to translate to public health systems.

Recent efforts via the Clinical Pharmacogenomics Implementation Consortium (CPIC) have aimed to standardise the nomenclature and terms for PGx test results (Caudle *et al.*, 2017). These terms relate the allele, and the allele combination (genotype) for variants in pharmacogenes to the phenotypes which occur as a result. These defined phenotypes are useful to assess risk for patients and to guide treatment based on the gene-drug combination. For all genes, allele functional status is defined into the following phenotypic categories: "increased function" (for function greater than normal), "normal function" (for the fully functional wild-type), "decreased function" (for those that function less than normal) and "no function" (for a complete non-functional state). "Uncertain function" is used to describe cases where literature supporting allele function is weak or conflicting, and "unknown function" is used when there is no literature available or when the variant is novel. Defined phenotypes are more specific to the gene category related to drug function, and relate the relationship based on the gene's mechanism of action with regard to the drug. Gene categories are divided into three main groups: drug metabolising enzymes, drug transporters, and genes which form high risk genotype status.

For drug metabolising enzymes, the phenotypes (Caudle *et al.*, 2017) are described based on the number of effect alleles present in an individual which directly impact enzyme activity. Individuals with two alleles that lead to increased function (as compared to wild type), or more than two normal (wild-type) alleles (this is possible through copy number variants for specific genes), are defined as "ultra-rapid metaboliser". "Rapid metaboliser" is defined via of a combination of increased function and normal function alleles. Increased function alleles are most commonly due to changes in copy number of the entire gene region. "Normal metaboliser" is defined by a combination of normal and decreased function alleles, but this phenotype is dependent on the nature of the decreased function allele. "Intermediate metabolisers" range between normal and decreased enzyme activity and are defined as combinations of normal, decreased function and/or function alleles. These relationships are displayed graphically in Figure 1.1 below to clarify that combinations of normal/decreased can fit either phenotype. The nature of the decreased function allele, in the case of missense variants, is related to the impact of the variant on protein structure. Some missense variants will have smaller changes to structure, and result in a functional enzyme,

albeit with reduced activity rates, whereas others that significantly alter structure or the active site can severely reduce or knock out enzyme activity. “Poor metaboliser” is defined by a combination of decreased functional alleles and or no function alleles (Caudle *et al.*, 2017).

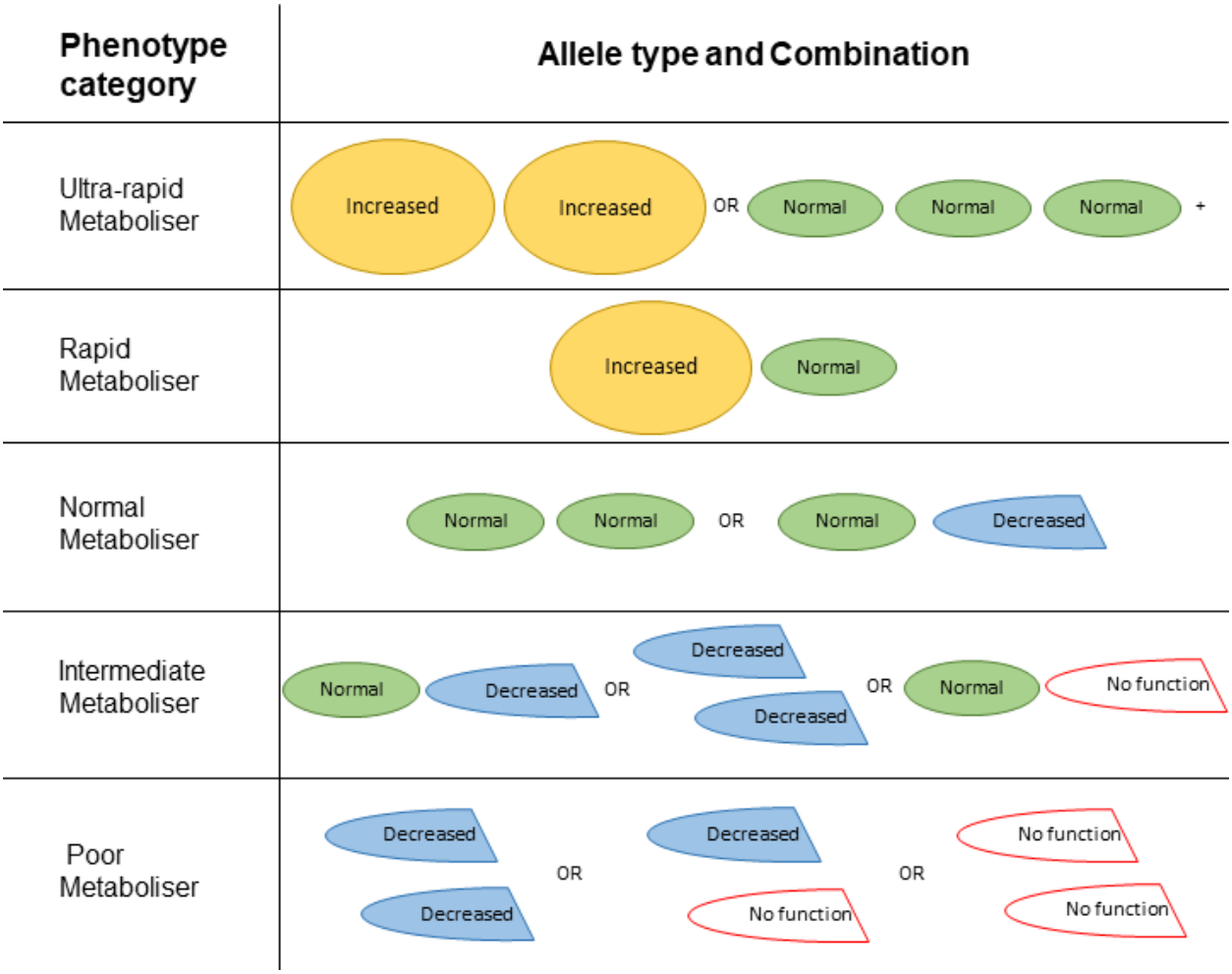


Figure 1. 1 Pictogram of allelic combinations (i.e. genotypes in individuals) that define phenotypes for drug metabolising enzymes (adapted via descriptions from Caudle *et al.* 2017)

The genotype combinations of variants can lead to multiple phenotypes, most notably there is overlap in the combination that can form normal or intermediate metaboliser status. Note that a combination of decreased function alleles can also form a poor metaboliser phenotype, but this is dependent on the level of enzyme activity drop that the variant induces. It must be noted that while the phenotype can be clearly defined, the effect on the patient, whether toxicity, efficacy or otherwise, remains highly dependent on the drug itself. A poor metaboliser gene status may mean poor efficacy in the case of one drug but may lead to severe toxic effects of another. Ultra-rapid

and rapid metabolisers may need increased dosages of some drug types but can result in toxicity effects due to high metabolite concentrations of other drug types (Caudle *et al.*, 2017).

Variants within genes encoding metabolic enzymes are commonly grouped into haplotypes which are associated with drug response. These groupings are referred to as star (*) alleles, and are used as a convention to describe the relationship between variant series and expected changes in drug pharmacokinetics and pharmacodynamics (Kalman *et al.*, 2016). Though all DNA variants in a particular haplotype (star grouping) are used to refer to the effect, this does not necessarily indicate that each variant alone has functional impact, and the effect may only be linked to the functional variant. Some star alleles are characterised by single variants of functional relevance making them easier to detect with a single SNP assay.

For drug transporters, the phenotypes relate the rate of transfer of drug with the allelic combination of variants (Caudle *et al.*, 2017). The definitions for these combinations resemble those defined for drug metabolising enzymes, with the exception that "rapid types" are not used. "Increased function" is defined via one or more increased functional allele, which leads to increased transporter function compared to normal. "Normal function" is defined as a combination of normal and/or decreased function alleles, while "decreased function" is defined via a combination of normal, decreased function and/or no function alleles. Poor function is defined via a combination of decreased function and/or no function alleles, which results in little or no transporter function . The clinical effect of these transporter phenotypes on the patient (efficacy etc.) is also highly dependent on the drug itself .

High risk genotype status, describes a phenotype specific to the gene-drug relationship. These genes encode molecules that are not necessarily directly involved in the metabolism or transport of a drug, but that are involved with other aspects or mechanisms of drug action. An example would be *HLA-B,* within which variants have been found linked to severe adverse reactions to allopurinol, a commonly used treatment for gout (Singh, 2013). *HLA-B* variants have been found to significantly increase the risk of a toxic reaction to allopurinol, though the gene is related to neither metabolism nor transport of the drug. These high-risk genotypes may also be the targets of drugs, for which changes can result in the drug failing to complete its mechanism of action (Caudle *et al.*, 2017). The phenotypes for these high-risk genes are defined into two categories "positive

risk" and "negative risk". The "positive risk" is defined by one or more high risk allele, and "negative risk" is defined by the absence of any high-risk allele.

PGx has been developing at a fast rate in recent years due to the advantages it provides in terms of information on drug dosage and drug safety. The phenotype definitions described above were designed to facilitate future work in the field and help reduce ambiguity in clinical reports, but these can also be used to describe the potential risk for patients based on their genotype for known PGx variants. By evaluating alleles, (and the variants that comprise them) and linking their known effect in other populations – PGx risk factors in populations that are currently under-characterised clinically can be evaluated if there is available genomic data. To date, the majority of PGx data have been collected based on research on individuals of European and Asian ancestry, whereas very little is known about continental African populations (Warnich *et al.*, 2011). This is particularly relevant for cancer treatment in Africa, as the majority of FDA recommendations for ethnicity related effects are for cancer drugs. As the rates of cancer are rising in Africa (while taking into consideration the difficulties faced by African patients), the need for pharmacogenomic research relevant to cancer and cancer drugs in continental Africans is highly emphasised. Attaining pharmacokinetic data for cancer drugs is not always feasible, due to the low therapeutic index of these drug types (Lee *et al.*, 2005). Thus, new strategies are needed to study PGx to address the emerging situation.

1.2 African Genetic Datasets and PGx data mining potential

The advent of next generation sequencing (NGS) allowed genetic research to significantly increase in output and has led to numerous advances in understanding cancer epidemiology, as well as the PGx of cancer treatments. Most PGx studies to date have taken place in non-African populations, but the recent release of African whole genome sequence data provides the opportunity to mine pharmacogenomically relevant data. With the completion of two major genetic projects, the largest amount of whole genome sequencing (WGS) data to date has been generated for African populations, forming two large, freely available datasets.

The first is the 1000 Genomes Project (KGP) (The 1000 Genomes Project Consortium, 2015), which produced WGS data for 26 global populations. Of these, seven were from African origins: the Yoruba (YRI) and Esan (ESN) from Nigeria; the Mende (MSL) in Sierra Leone; the Gambian (GWD) in Western Divisions in the Gambia; the Luhya (LWK) in Webuye, Kenya; Americans of

African Ancestry (ASW) in the South Western United States; and African Caribbean's (ACB) in Barbados. The KGP ran over three phases: I, II, and III. Phase I was a trial phase of low coverage WGS and exome sequencing, on 1092 samples. Phase II expanded the study to 1700 samples and included new methods to call multiple base variant sites (insertions or deletions). Phase III included the total 2504 samples, including the African samples. Sequencing was conducted at a mean coverage of 4x across all samples. Coverage refers to the number of times a single nucleotide is read during the sequencing process, and can represent the relative confidence of variant discovery (Sims *et al.*, 2014).

The second set is from the African Genome Variation Project (Gurdasani *et al.*, 2015) which has significantly characterised African genetic variation for populations spanning the continent, and created a large resource to facilitate biomedical studies in Africa. This study generated SNP array genotype data from 1481 individuals from 18 ethno-linguistic groups, and WGS data (Figure 2.1) from 320 individuals from Ethiopia, South Africa, and Uganda. The South African data included 100 Zulu individuals with approximately 4-fold coverage, the Ugandan data included 100 Bagandan individuals (the most common ethnic group in the country) with approximately 4-fold coverage and the Ethiopian data included 120 individuals from three main ethnolinguistic groups with a 4-8-fold range of coverage.

The WGS datasets from the KGP and the AGVP studies are available in VCF file format. The VCF file is produced from an alignment file, and is a representation of the variants present in the sequenced genome compared to a reference genome (both AGVP and KGP used GRCh37 as a reference), and contains information related to the nature and quality of the call of the variants such as: the genotype at the variant position, quality scores, and the read depth (how many overlapping contiguous sequences include that variant position) (Standish *et al.*, 2015). This file format is ideal for large scale variant analysis, as it is compact compared to alignment files, and supported by multiple bioinformatics tools.

The WGS data for the African genomes can be used to identify variants in African populations and these variants can be assessed to predict their potential functional impact. A number of annotation tools exist to determine the predicted functional effect of a variant. For example, a loss of function variant, or the effect of amino acid substitutions from missense variants on relevant protein functions and how they may affect a therapeutic pathway. Ensembl's Variant Effect

Predictor (VEP) (McLaren *et al.*, 2016) is a customisable tool that can predict the effect of variants such as SNVs, deletions, insertions, and structural variants on genes, their corresponding transcripts and proteins. It is a powerful variant annotation tool that has access to Ensembl's repositories of genomic data. It is also highly customisable to include additional software to provide layers of annotation for all variant types. In lieu of clinical trials, which are costly and difficult to motivate in resource poor environments, bioinformatics tools can be used to mine pharmacogenomically relevant data from African whole genome sequence data.

1.3 Background on genes relevant to cancer related pharmacogenomics

In this section I provide detailed information on thirteen genes of interest. They were selected in line with the objectives but need to be described in the introduction as part of the background literature to the study. To characterise overall possible drug-gene effects in cancer treatment, these thirteen genes were selected based on the most common cancer types in Africa, the number of drugs that have adverse reactions in these types, and the primary role of the gene in the drug pathway. The genes selected encompass four main roles: encoding key enzymes needed in the metabolism of drugs, encoding drug transporters, encoding the molecular targets of drugs, and genes that encode specific DNA repair proteins (e.g. for platinum-based cancer drugs) (summarised in Fig. 1.2). Disruption of these gene-roles may have various outcomes following or during treatment, such as toxic effects related to drug concentration, or lowered efficacy of treatments, which decreases the chance of survival. Details regarding the location and size of the transcript and protein are listed in Table 1.2 below.

Table 1. 2 Location and size data for genes analysed (GRCh37).

Gene Name	Location (chr:position; Genome Reference 37)	Canonical Transcript Location (chr:position; Genome Reference 37)	Canonical Transcript Length – Exons and UTRs (bp)	Protein Length (aa)	Direction of Transcription	Primary Role in respect to drug
ABCB1	7: 87,132,175-87,343,611	7: 87,133,175-87,342,564	4645	1280	Reverse	Transporter
CYP19A1	15: 51,499,254-51,631,807	15: 51,500,254-51,630,807	4417	503	Reverse	Metaboliser
CYP1B1	2: 38,293,116-38,338,044	2: 38,294,652-38,303,323	5247	543	Reverse	Metaboliser
CYP2D6	22: 42,521,501-42,527,908	22: 42,522,501-42,526,908	1684	497	Reverse	Metaboliser
CYP3A4	7: 99,353,604-99,382,888	7: 99,355,299-99,381,888	2153	503	Reverse	Metaboliser
CYP3A5	7: 99,244,817-99,278,621	7: 99,245,817-99,277,619	1720	502	Reverse	Metaboliser
DPYD	1: 97,542,299-98,387,605	1: 97,543,299-98,386,579	4412	1025	Reverse	Metaboliser
ESR1	6: 151,976,826-152,451,754	6: 152,128,686-152,424,406	6455	595	Forward	Target
GSTP1	11: 67,350,066-67,355,131	11: 67,351,066-67,354,131	965	210	Forward	Metaboliser
SLC19A1	21: 46,912,486-46,965,325	21: 46,933,690-46,962,385	3811	591	Reverse	Transporter
TYMS	18: 656,604-674,578	18: 657,604-673,578	1662	313	Forward	Target
XRCC1	19: 44,046,192-44,085,625	19: 44,047,192-44,080,158	2802	633	Reverse	DNA Repair
XRCC5	2: 216,971,187-217,072,026	2: 216,972,187-217,071,026	3761	732	Forward	DNA Repair

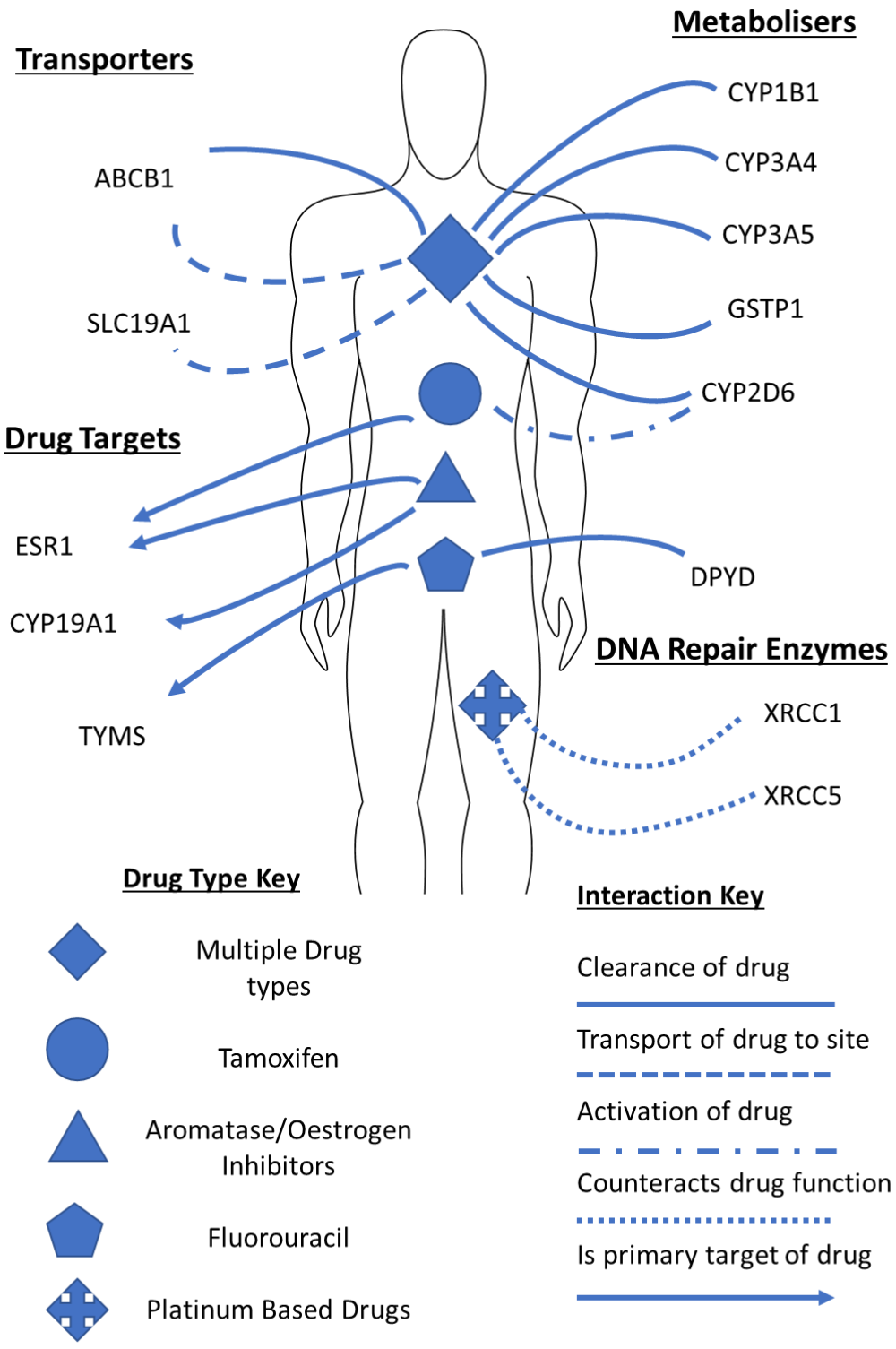


Figure 1. 2 Drug-Gene Interaction Summary Plot for genes analysed

This figure highlights non-exhaustive broad relationships and interactions between the genes selected for analysis and key drugs with which they interact. Location of markers has no relevance to drug administration location.

1.3.1 Drug Transporters

These are genes involved with the movement of drugs across membranes, facilitating their movement across cells. These can be involved from the point of administration to the site of drug action. This section describes the genes selected that fall into the category of drug transporters and highlights their principle PGx related effects.

ATP-Binding Cassette sub-family B member 1 - ABCB1

The *ABCB1* gene, encodes a membrane protein called P-glycoprotein 1, which actively transports xenobiotics across cell membranes (Bodor *et al.*, 2005). It is expressed within the plasma membrane, and acts as an efflux pump eliminating drugs from cells. It plays a critical role in elimination of drugs during first pass elimination by pumping them into the excretory systems of the liver and small intestines, lowering their overall bioavailability. This protein is also involved in protection of key organ systems from xenobiotics, particularly in the blood brain barrier. When upregulated, this protective function, however, is deleterious to the function of psychiatric drugs targeting the brain, and genotypes of *ABCB1* have been associated with treatment failure with antidepressants and other psychoactive drugs (O'Brien *et al.*, 2012). It is also found overexpressed in cancer cells exhibiting strong resistance to treatment, as the drugs are pumped out before reaching concentration levels needed to produce the therapeutic effect (Mahon *et al.*, 2003). Due to its role in reducing the therapeutic effect of multiple drug types this gene has also been referred to as the Multi-Drug Resistance Protein 1.

This gene has historically been well characterised in terms of its pharmacogenomic effect in cancers for drug resistance, usually as a result of overexpression on the cell surface of cancer cells, but more recent investigations have examined the role of germline coding region variants on overall drug response (Yao *et al.*, 2014), including toxicity related effects as well as efficacy. Variants which are damaging to *ABCB1* function could result in improved efficacy of drugs, as there would be increased bioavailability, but this in turn could result in toxic effects, as the drug would not be cleared from the body before producing toxic effects. This has been seen particularly with the drug, methotrexate (Zgheib *et al.*, 2014), which is commonly used to treat multiple cancer types. This genes effect thus can be manifested in two ways, through quantitative change in gene expression, and through allelic variation that alters its protein function. Variants that can influence either of these effects would be useful for PGx strategies regarding *ABCB1*.

Solute Carrier Family 19 Member 1- SLC19A1

The *SLC19A1* gene encodes the Reduced Folate Carrier 1 protein (RFC1), which is the major system for folate transport in cells and tissues. Folates form part the family of B vitamins and require active transport to cross cellular membranes. Folate is a key nutrient needed to facilitate cell division, DNA synthesis and enzyme substrate methylation. The gene is expressed ubiquitously, as it is the primary route of folate into cells (Zhao and Goldman, 2013). Cancer cells which grow rapidly and divide constantly require increased nutrient uptake to support their expansion, and are able to increase the activity of nutrient uptake pathways via folate-dependant synthesis of nucleotides (Gonen and Assaraf, 2012). Drugs targeting the folate metabolism system are used as chemotherapies which slow the expansion of cancerous cells. RFC1 is expressed ubiquitously, and is well characterised for its role in resistance to anti-folate based chemotherapies, such as methotrexate (Huang, 2007).

Numerous single nucleotide variants (SNVs) within the *SLC19A1* gene region have been found associated with decreased treatment response to methotrexate (Liu *et al.*, 2017; Owen *et al.*, 2013), but most current findings are for European or Asian populations. The reduced treatment response can lead to increased chance of recurrence, shorter survival times, and additional treatments and costs to the patient.

1.3.2 Drug Metabolisers

Metaboliser genes encode proteins that are involved in the breakdown of drug molecules, either for activation of the drug or for removal from the body. This section describes the genes selected that fall into the category of drug metabolisers, highlighting key PGx related effects regarding efficacy or toxicity of the drug based on genetic variation.

The Cytochrome P450 Super-Family

The cytochrome P450 proteins are a super family of metabolic enzymes. These enzymes are numerous in both their type and the number of substrates with which they interact, which include cholesterols, steroids and other lipids, as well as a variety of drug types. Members of this protein family are characterised by monooxygenase activity which enables their metabolic function, and are mostly localised to the endoplasmic reticulum. These molecules are large in size and flexible to accommodate a broad range of substrates, and are responsible for hepatic clearance of 78% of drugs currently in use (Zanger *et al.*, 2008).

Cytochrome P450 Family 1 Subfamily B Member 1 - CYP1B1

The *CYP1B1* gene encodes a widely expressed enzyme that catalyses a variety of substrates via oxygenation reactions. It is localised to mitochondrial membranes, where it is a key regulator of steroid, lipid and vitamin homeostasis. It also metabolises a broad spectrum of xenobiotic compounds including numerous cancer drugs. Variants within the gene have been found to compromise enzymatic activity, lowering activity of its metabolic processes by up to 50% (Jansson *et al.*, 2001). Lowered activity may lead to slower clearance of drugs. Variants within this gene have been found associated with increased risk of toxic adverse effects following treatment with cyclophosphamides, paclitaxel and epirubicin (Abraham *et al.*, 2014), which are all chemotherapies used to treat breast cancer, often in combination.

Cytochrome P450 Family 2 Subfamily D Member 6 - CYP2D6

The *CYP2D6* gene also has a wide variety of substrates, and is extremely well characterised in terms of pharmacogenomic effect, as the majority of psychiatric and antidepressants are metabolised by this enzyme (Winner *et al.*, 2013).

CYP2D6 genotyping is particularly relevant for oestrogen receptor positive breast cancer patients, who are commonly treated with tamoxifen. Tamoxifen is a prodrug, which must be activated via CYP2D6 metabolism into its active form, Endoxifen, which then regulates oestrogen pathways to lower to risk of breast cancer recurrence. Variants in the gene have been linked to increased chance of cancer recurrence, as well as poorer survival outcomes for the patient (Saladores *et al.*, 2015).

Adverse reactions and treatment failure for numerous drugs, including Tamoxifen, have been associated with SNVs in the gene itself that reduce enzyme function (Ingelman-Sundberg, 2005), variants that can induce alternative gene splicing (Wang *et al.*, 2014), and copy number variants of the entire gene (Ledesma and Agundez, 2005). SNVs found linked to adverse drug reactions are more common than other variant types, though both these and other types are included in genotyping assays for *CYP2D6* that aim to guide treatment (Rebsamen *et al.*, 2009). Though these assays are commonly used, the genotyping assays (which are based mostly on European and Asian variants) may underperform in African cohorts (Dodgen *et al.*, 2013).

Cytochrome P450 Family 3 Subfamily A Member 4 - CYP3A4

The *CYP3A4* gene encodes a multi-functional enzyme which is responsible for metabolism of 60% of all drugs metabolised by the cytochrome family (Preissner *et al.*, 2013). It is a flexible enzyme,

accommodating a broad range of substrates, and often more than one at the same time. It is expressed in most body tissues but predominantly in liver (Ohtsuki *et al.*, 2012), where it plays key roles in both drug activation and clearance. Due to its broad specificity, variation within the gene can have an impact on a wide variety of drug types.

Variants within the *CYP3A4* gene have been associated with toxicity of cancer drugs such as docetaxel (Boso *et al.*, 2014), paclitaxel (Nakajima *et al.*, 2006) and with increased risk of cancer following tamoxifen treatment (Chu *et al.*, 2007). However, these findings are in European and Asian based cohorts with little data on Africans. *CYP3A4* variants are known to affect efavirenz (antiretroviral) treatment in African populations (Haas *et al.*, 2004), though it is still not clear if these variants will affect cancer treatments in the same manner in these populations.

Cytochrome P450 Family 3 Subfamily A Member 5 - CYP3A5

The *CYP3A5* gene, along with *CYP3A4*, are the most predominantly expressed of the cytochrome P450 family in the liver. The 3A5 enzyme is similar in structure to that of 3A4 and these two genes product overlap in most of their respective substrates.

This gene has been well characterised for its role in tacrolimus metabolism. Tacrolimus is a drug used to prevent organ rejection following a transplant, where high doses can prove toxic, but low concentrations fail to prevent transplant rejection (Passey *et al.*, 2011). Variants in 3A5, including the 3A5*3 allele, which is a knockout inducing allele, have been found linked to tacrolimus response (Quteineh *et al.*, 2008). The wild type allele has been linked to increased risk of organ rejection compared to individuals with the 3A5*3 variant due to reduced concentrations of tacrolimus (Borobia *et al.*, 2009). The 3A5*3 variant is extremely common in Europeans and is near fixation (Valente *et al.*, 2015), but is uncommon in African populations where *CYP3A5* variation is more diverse (Rajman *et al.*, 2017).

While this knockout variant is beneficial to tacrolimus treatment outcomes, it may induce toxicity of others, including key cancer drugs. Low expression of *CYP3A5* has been found associated with vincristine induced neurotoxicity in children with acute lymphoblastic leukemia (Egbelakin *et al.*, 2011). The 3A5*3 variant has also been associated with toxicity of paclitaxel and carboplatin drugs in Asian ovarian cancer patients (Hu *et al.*, 2016), and with paclitaxel induced neurotoxicity in Spanish cancer patients (Leskela *et al.*, 2011).

Glutathione S-Transferase Pi 1 - GSTP1

The *GSTP1* gene is in the family of glutathione S-transferase enzymes, which play important roles in metabolism of toxic compounds via conjugation with reduced glutathione (Townsend and Tew, 2003). The pi classes of these enzymes are cytosolic and play regulatory roles in the mitogen activated kinase pathways, which co-ordinate cellular survival. It the smallest gene compared to the others selected in this analysis (Table 1.2). *GSTP1* is widely expressed throughout human tissues and plays a central role in cellular detoxification. Some cancer types are resistant to treatments due to overexpression of GSTP1, which improves cellular detoxification efficiency (Depeille *et al.*, 2005).

GSTP1 is well characterised for its role in cancer drug metabolism (Sawers *et al.*, 2014; Tulsyan *et al.*, 2013), and is included in genotyping strategies to be used prior to treatment with fluoropyrimidine and oxaliplatin based therapies (Di Francia *et al.*, 2015) to aid in successful treatment outcomes. A missense variant within the *GSTP1* gene, rs1695, has been found linked to increased risk of toxicity following treatment with platinum based drugs (Kim *et al.*, 2009) in patients carrying the A allele.

The same rs1695 allele however, has been found linked to increased drug response to cyclophosphamide and epirubicin treatments in breast cancer patients (Oliveira *et al.*, 2010; Zhang *et al.*, 2011) as well as with reduced severity of toxic effects of these drugs in Japanese patients with early breast cancer (Sugishita *et al.*, 2016). This indicates multiple pharmacogenomic actions of the same gene, whereby *GSTP1* variants may increase likelihood of treatment failure and increase risk of drug toxicity for some drug types, while causing other drug types to be more effective depending on the alleles present in the patient.

1.3.3 Specific Drug Targets

These are molecules that are the targeted sites for drug binding and interaction in order to conduct their therapeutic mechanisms. This section describes selected genes that fall into the category of drug targets, highlighting PGx related effects due to changes in the target or its expression.

Oestrogen Receptor 1 - ESR1

The *ESR1* gene encodes the human oestrogen receptor alpha protein, a key transcription factor activated via oestrogen hormone (Dahlman-Wright *et al.*, 2006). Oestrogen and oestrogen receptors are key for sexual development and reproduction processes, but also are active in other

tissue types. Oestrogen receptor alpha is expressed in the nucleus, where it may form homodimers, or heterodimers with oestrogen receptor beta. It has multiple domains which bind its substrate, as well as to DNA, enabling it to act as a transcription factor.

As part of the oestrogen signalling system, this molecule is targeted via oestrogen receptor agonists to inhibit signalling pathways and slow cancer growth. Variants within *ESR1* can cause resistance to hormone based therapies (Li *et al.*, 2013), and can result in higher chances of experiencing musculoskeletal pain during treatment (Wang *et al.*, 2013). Tamoxifen treatment is known to help lower serum cholesterol in women during treatment, however, these changes can be affected by *ESR1* variants, which may affect the level of decreased total cholesterol in postmenopausal women (Ntukidem *et al.*, 2008). These findings, however, are mostly reported from cohorts of European and Asian women. Oestrogen receptor positive breast cancers are predominant in sub-Saharan Africa (Eng *et al.*, 2014), which urges further analysis into drugs targeting oestrogen pathways.

Human Aromatase - CYP19A1

The *CYP19A1* gene encodes for human aromatase enzyme, (commonly referred to as oestrogen synthetase), and is a key part of the process of producing oestrogens from catalysis of androgens. It is also a member of the cytochrome superfamily but is specific to its androgenic substrates: androstenedione and testosterone, which it transforms into oestrone and oestradiol respectively. Oestrogens are steroid hormones which regulate a variety of biological processes, particularly in women, via hormone signalling mechanisms. These signals bind to oestrogen receptors, which in turn control gene expression in their respective cells. These signalling pathways are often involved in hormone regulated breast cancers, in which oestrogen pathways are upregulated to improve cancer cell growth (Samavat and Kurzer, 2015). Mutations in the gene which are highly damaging to gene function can lead to altering aromatase function, which may cause aromatase deficiency leading to pseudo-hermaphroditism in females (Jones *et al.*, 2007).

This enzyme is the target of aromatase inhibitors, which are drugs used to prevent the production of oestrogen, and are primarily used to treat hormone-receptor positive breast cancer by inhibiting growth signals mediated by the hormone (Baselga *et al.*, 2012). Other variants which may produce less severe structural changes or changes in gene expression have been found linked to reduced efficacy of treatment, which can lead to cancer recurrence (Colomer et al., 2008; Shao et al., 2015a).

1.3.4 Fluoropyrimidine-related pharmacogenes

This section details genes that interact with the same group of drugs, fluoropyrimidines, which are widely used as chemotherapeutics. The two genes consist of the primary metaboliser, dihydropyrimidine dehydrogenase, and the intended drug target, thymidylate synthase.

Dihydropyrimidine Dehydrogenase - DPYD

The *DPYD* gene encodes dihydropyrimidine dehydrogenase, an enzyme involved in the metabolism of fluoropyrimidines, such as fluorouracil (a WHO essential medicine), capecitabine and tegafur. Fluoropyrimidines are commonly used to treat a wide variety of cancers and exert their anti-cancer effect in multiple ways, but primarily via inhibition of the thymidylate synthase enzyme. Thymidylate synthase plays a key role in DNA replication (see below). The *DPYD* gene is expressed widely throughout the body, but highest expression is found in liver and peripheral blood.

DPYD variants that lower expression or produce non-active proteins can cause severe, life threatening side effects following standard doses of fluoropyrimidine based drugs (Loganayagam *et al.*, 2013). DPYD partial deficiency and complete deficiency are present approximately 4% and 0.2% of Caucasians respectively (Morel *et al.*, 2006). SNVs within the *DPYD* gene have been found linked to increased risk of severe toxicity following fluorouracil treatment in European (Saif, 2013) and Asian (Zhang *et al.*, 2012) cancer patient cohorts, and a variant specific to African Americans (compared to European Americans) has been found to reduce DPYD enzyme activity (Offer *et al.*, 2013). Genotyping strategies are currently being developed to generate an assay that identifies at risk patients, which can be then used to prevent toxic events and ensure effective treatment for the patient (Lunenburg *et al.*, 2016). These work by identifying at risk patients, which can then have dosage adjusted or an alternative treatment given. Although useful, current strategies are guided by variants identified in studies of mainly Europeans and Asian populations and may not be useful for African patients at the present time.

Thymidylate Synthase - TYMS

Thymidylate synthase, encoded via *TYMS*, is an enzyme that catalyses the formation of deoxythymine monophosphate (the thymine base of monomeric DNA) from deoxyurimidine monophosphate. Thymidylate synthase plays key roles in the folate-homocysteine cycle (specific amino acid synthesis) and in purine and pyrimidine synthesis (Wilson *et al.*, 2014). As it is a

primary producer of thymine deoxynucleotides, this enzyme is key to DNA synthesis, replication and repair.

The enzyme has been well characterised as a chemotherapeutic target (Danenberg *et al.*, 1999), as inactivation of its pathways will rapidly starve fast replicating cells of thymine. Fluoropyrimidines are most commonly used as antimetabolites targeting the TYMS enzyme, and inhibit its activity via irreversible binding (Papamichael, 1999). Fluorouracil is a widely used fluoropyrimidine based drug, and is a WHO essential medicine, meaning it is intended to be available at functional healthcare systems in adequate dosage and at affordable prices to ensure the public has access to the drug. These drugs are extremely toxic, and often induce severe adverse reactions. Adverse reaction risk to fluorouracil treatment is most commonly associated with *DPYD* variants (as the active metaboliser of these drug types), but variants that alter levels of active TYMS enzyme have also been found linked to severe drug side effects (Lima *et al.*, 2013).

A six base pair deletion in the 3-prime untranslated region of *TYMS* has been found to improve survival time post treatment but increase risk of toxicity in a mixed population cohort of colorectal cancer patients (Rosmarin *et al.*, 2015). Patients homozygous or heterozygous for the allele have reduced survival times, but lower risk of toxicity during treatment. Other variants which are SNVs have been linked to increased risk of drug toxicity following fluoropyrimidine based treatments (Meulendijks *et al.*, 2016; Pellicer *et al.*, 2017).

1.3.5 DNA Repair Proteins

DNA repair genes encode proteins that bind to DNA repair enzymes and co-ordinate them to repair damaged strands. These are relevant to cancer drugs that aim to damage cancer cell DNA in order to destroy them or slow their growth. This section describes the genes selected which have the primary role of DNA repair and are involved in specific drug interactions related to this process.

X-Ray Repair Cross Complementing 1 - XRCC1

The *XRCC1* gene encodes a DNA repair protein scaffold, X-ray repair complementing protein 1, which binds to single stranded DNA, along with DNA ligase III and other DNA repair enzymes to co-ordinate them for efficient repair of signal and double strand breaks, base excisions, and nucleotide excisions. This gene is overexpressed in non-small cell lung carcinoma, which is unusual as most cancer types are deficient for one or more DNA repair proteins or enzymes. A

truncated form of the *XRCC1* protein has been found to suppress tumour growth in multiple cancer types (Pettan-Brewer *et al.*, 2012).

Mutations in this gene have been found to affect platinum based drug efficacy (Yuan *et al.*, 2015), and can be used as predictive markers to guide treatment. Platinum based drugs, such as cisplatin, induce crosslink formations in DNA (Rudd *et al.*, 1995), which are covalent bonds between bases. These crosslinks form in an inter-strand or intra-strand manner, and interfere with DNA replication and repair mechanisms, slowing the growth of cancer cells. *XRRC1* variants that disrupt gene function have been shown to affect response to platinum drugs in Asian cervical cancer patients (Chung *et al.*, 2006), and affect the survival of Asian colorectal cancer patients following treatment with fluorouracil, leucovorin and oxaliplatin (Huang *et al.*, 2011). The *XRCC1* gene is likely protecting cancer cells from the treatments by correction of the crosslinked DNA, allowing the cells to grow freely.

X-Ray Repair Cross Complementing 5 - XRCC5

The X-Ray Repair Cross Complementing 5 protein, also known as Ku80 is encoded by *XRCC5*. Ku80, along with Ku70 (encoded by XRCC6) forms a heterodimeric repair protein called Ku, which binds to double stranded DNA breaks and mediates the non-homologous end joining pathway of DNA repair (Walker *et al.*, 2001). Like *XRCC1*, this gene may contribute to cancer cells resisting treatment via repair of drug induced DNA damage (Shang *et al.*, 2017). Reduced expression of XRCC5 may cause patients to respond better to drugs targeting cancer cell DNA.

XRCC5 variants have been found to increase overall survival in Spanish multiple myeloma patients treated with thalidomide (Cibeira *et al.*, 2011), as well as with reduced toxicity to platinum based drugs in Asian cohort of lung cancer patients (Chen *et al.*, 2016).

1.4 Study Rationale

There is a scarcity of information on cancer related PGx in African populations. Ideally, mass clinical trials and pharmacokinetic assays would be necessary to address the situation, but these are costly and difficult to co-ordinate in low resource environments. With large publicly available datasets of African whole genome data, it is possible to mine variants that are potentially relevant to cancer PGx. Cancer types which are most prevalent in Africa and for which mortality is predicted to increase drastically were selected to evaluate both the most commonly used treatments (drugs) and the genes that are involved in their action and efficacy.

This promoted the present study that started by developing a list of cancer relevant PGx genes to assess African cancer PGx. Only public WGS data were used to extract genetic variants in these genes and to use bioinformatic *in silico* approaches to characterise variants identified from continental African populations. This could lead to the identification of novel variants that may influence treatments and the data could also be used to compare the frequencies of known PGx variants between other populations and Africans.

By finding rare and common variants likely to be deleterious, this study could guide future work and functional assessment to characterise their clinical relevance. Although rare variants are less likely to impact many people, they could be critical for the individual and could also serve to provide deeper insights into gene-drug interactions. Common variants predicted to have significant functional consequence could impact many patients receiving treatment and could be used to motivate clinical trials and studies to validate clinical impact, and to provide insights into policies and strategies for public precision medicine in Africa.

1.5 Aim and Objectives

The broad aim of this study was to assess and characterise cancer PGx relevant variants in continental Africans and to compare the frequency of novel and known variants to other global populations.

Hypothesis

WGS data will reveal novel PGx variants and comparing allele frequencies for novel and known PGx variants will show differences in continental African populations when compared to European, Asian and American populations.

Aim

To characterise and annotate genetic variation in selected cancer PGx related genes in African populations using whole genome sequence data from publicly available genome data.

Objectives

1. To identify relevant cancer types and create a list of genes that are relevant to pharmacogenomics of cancer drugs in Africa.
2. To mine public WGS data for novel and known PGx variants in the 1000 Genomes Project and the African Genome Variant Project datasets for African populations.
3. To perform and assess functional annotation on the variants and rank them in order of their potential risk of involvement in adverse drug reactions.
4. To determine frequencies of relevant variants in African populations and compare them to those observed in European, Asian and American populations.

Chapter 2: Methodology

2.1 Genetic Datasets

Two large publicly available WGS data sets with significant African data were used to conduct a bioinformatics analysis of variants that may be involved in pharmacogenomics-based adverse reactions with cancer drugs; and to compare the frequencies of these variants in Africans to other global populations.

WGS data for the Yoruba, Esan, Mende, Gambian and Luhya populations was selected for use from the KGP datasets. These datasets are freely accessible at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data and variant call files were downloaded. Files were produced via a pipeline of bioinformatics tools, with multiple quality control steps in place to assess the accuracy of calls. Though numerous variant calling methods were applied, these mainly ensured quality by filtering out poorly mapped reads and poor quality read calls, aiming to keep a false discovery rate of below 5%. The KGP datasets have an average coverage of 7.4x (The 1000 Genomes Project Consortium, 2015) and the number of individuals sequenced for each population used in this study is shown in Figure 2.1. The exomic data generated in KGP Phase III was not used in this study. The three populations with WGS data from the African Genome Variation Project were also used. These populations are from Ethiopia, South Africa, and Uganda (Figure 2.1), with the prior having 4-8x coverage and the latter two having an average of 4x coverage (Gurdasani *et al.*, 2015). Variant call files for these three populations were obtained from the AGVP project which are accessible from the European Genome-phenome Archive (accession number EGAS00001000363). An ethical approval waiver for the use of these public datasets was obtained from the Human Research Ethics Committee (Medical), University of the Witwatersrand (Appendix 1).

The populations compared in this study (Figure 2.1) are defined as follows: “Continental Africans” include all African population datasets assessed. These were defined with the following nomenclature: “West Africans” include Yoruba (YRI), Esan (ESN), Mende (MSL) and Gambian (GWD); “East Africans” include Luhya (LWK), Bagandan (BAG) and Ethiopian (ETH); and “South Africans” include Zulu (ZULU). When frequencies are discussed regarding Africans, they refer to an average of the allele count for continental African populations (not including the ACB

and ASW KGP populations). The comparison populations, selected from the KGP, are the super-continental averages for Europeans, Admixed American, South Asians, and East Asians.

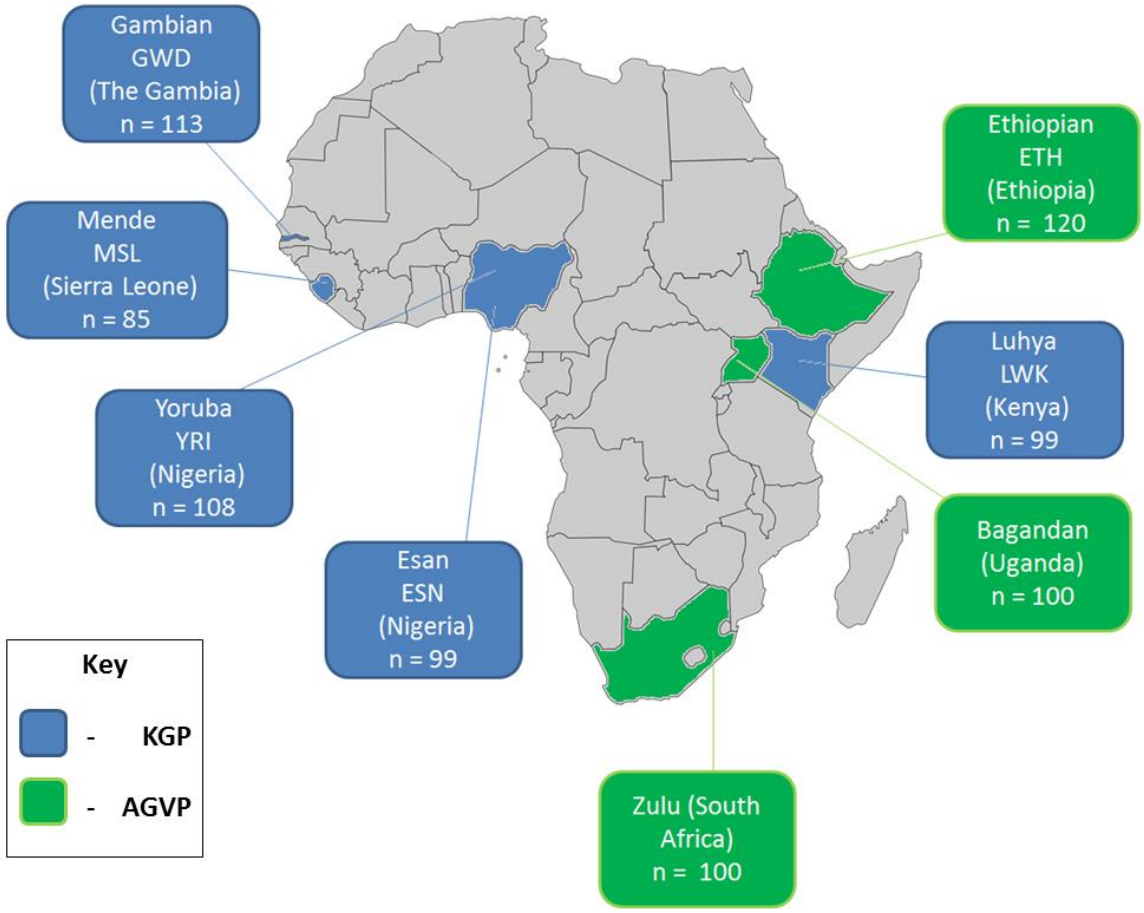


Figure 2. 1 Geographic origin of African populations used in this study of low coverage whole genome sequence data

This study used data only from the continental African populations to search for variants, to minimise effects of admixture and genetic drift on novel variant detection and allele frequencies in admixed populations. Therefore, the African American (ASW) and Afro-Caribbean (ACB) diaspora populations from the KGP were excluded as these populations have significant differences in substructure compared to continental Africans (Benn-Torres *et al.*, 2008; Bryc *et al.*, 2015; Murray *et al.*, 2010).

Sequencing coverage below eight fold will not have the capability to provide the complete genotype of a sample, but will allow for detection of most variants with frequencies of 1% or above in the groups studied (Gurdasani *et al.*, 2015). These datasets were refined by several quality control steps, so it can be accepted that the majority of variants detected are called with reasonable confidence. In this study, novel variants (refer to those variants not previously described, and that do not have an "rs" number in dbSNP) which are usually rare would require technical validation to ensure that they are not artefacts of NGS before they can be used for PGx studies. Common variants which are previously described, both in terms of sequencing and clinical function, are more reliable for PGx study use. This study defined common variants as those with a minor allele frequency (MAF) greater than or equal to 1%, and rare variants as those with a MAF less than 1%.

2.1.1 Selection of Cancer Types most relevant in Africa

The WHO data for cancer mortality (described above) shows that the most common cancers inducing mortality in Africa are: breast and cervical for women; lung, oesophageal and prostatic for men; and lymphoma/multiple myelomas and liver for both sexes. The mortality rates of these cancer types are also predicted to substantially increase by the year 2030. Due to how common they are and how likely these will occur in the future; these cancer types were selected to browse the PharmGKB database.

2.1.2 Selection of Genes to investigate

The PharmGKB database (accessed at: https://www.pharmgkb.org/ on the 20/04/2016) was mined for relevant drug-gene data for the selected cancer types including the following search terms: Oesophageal Neoplasms, Lung/Tracheal Neoplasms, Prostatic Neoplasms, Breast Neoplasms, Uterine-Cervical Neoplasms, Leukaemia, Multiple Myeloma, and Liver Neoplasms. Data extracted was specific to the cancer type, using the name as the search qualifier. All genes known to be linked to pharmacogenomic interactions within these cancer types were selected.

From these data, genes for this study were selected based on the number of cancer types they affected (preferably more than one), the number of drugs linked to the gene, and the severity of known adverse reactions to the drug. Effort was also made to include genes that are drug metabolisers, transporters, and those that can form high risk genotypes (e.g. drug targets). Data pertaining to these genes were extracted, including the cancer types in which they are involved, the known alleles associated with effects, and the drugs these alleles are associated with. The data

for all cancer types will be presented, with the exception of liver cancer where no data are available on PharmGKB. Lack of PGx data on liver cancer may be due to the biological function of the liver to filter and metabolise a number of drugs, and thus, any PGx phenotype detected may be confounded because of the malignancies of the liver tissue (Aboel Dahab *et al.*, 2016).

2.1.3 Selection of variants associated with the genes

Variants from the entire genomic sequence (exons and introns) for the genes that were selected, as well as a 1000 base pair (bp) flanking region (upstream and downstream of the gene) were extracted from the complete VCF files of each population dataset from the KGP and AGVP. This would be saved in eight smaller VCF files in total (one for each population) prepared for annotation, with each containing variants related to only the genes selected (although some overlap may occur with other co-located genes). The variants types include SNVs, as well as small insertions or deletions that have been called within gene coding and associated regions. All variant coordinates were mapped to the GRCh37 (hg19) version of the human genome assembly.

2.2 Functional Annotation

VEP (McLaren *et al.*, 2016), a powerful annotation tool available through the Ensembl genome browser, was used to annotate genomic variants with information regarding their type, functional impact and location. It works via comparing the variants used as input across a large genomic reference dataset. The GRCh37 reference genome sequence was used for this study. VEP can also incorporate other programs into its software to provide extra layers of information and prediction. These can be included as plugins, which are freely available at https://github.com/Ensembl/VEP_plugins. SIFT (Kumar *et al.*, 2009) and Polyphen-2 (Shihab *et al.*, 2013) are the most widely used tools for variant annotation in coding sequences (they do not provide annotation for non-coding sequences like 5' and 3'UTRs, introns and flanking regulatory sequences). These two programs are built into VEP, and thus do not need to be installed as additional plugins. For this study, VEP (release v86) was built up with a series of additional plugins that annotate variant data. The following plugins were included: Sorting Tolerant from Intolerant (SIFT) v5.5.2 (Ng and Henikoff, 2003); Polyphen-2 v2.2.2 (Adzhubei *et al.*, 2010); Combined Annotation scoRing toOL (CAROL) v1.0 (Lopes *et al.*, 2012); consensus deleteriousness score of missense mutations (Condel) v2.0 (González-Pérez and López-Bigas, 2011); Functional Analysis through Hidden Markov Models – MKL (FATHMM) v1.0 (Shihab *et al.*, 2015); Combined Annotation Dependent Depletion (CADD) v1.3 (Kircher *et al.*, 2014); and Loss-Of-Function

Transcript Effect Estimator (LOFTEE) v1.0 (https://github.com/konradjk/loftee). A flow diagram of how the VCF data was analysed with this toolset is shown in Figure 2.2 below, while the functions of the tools and plugins are described in the sections that follow.

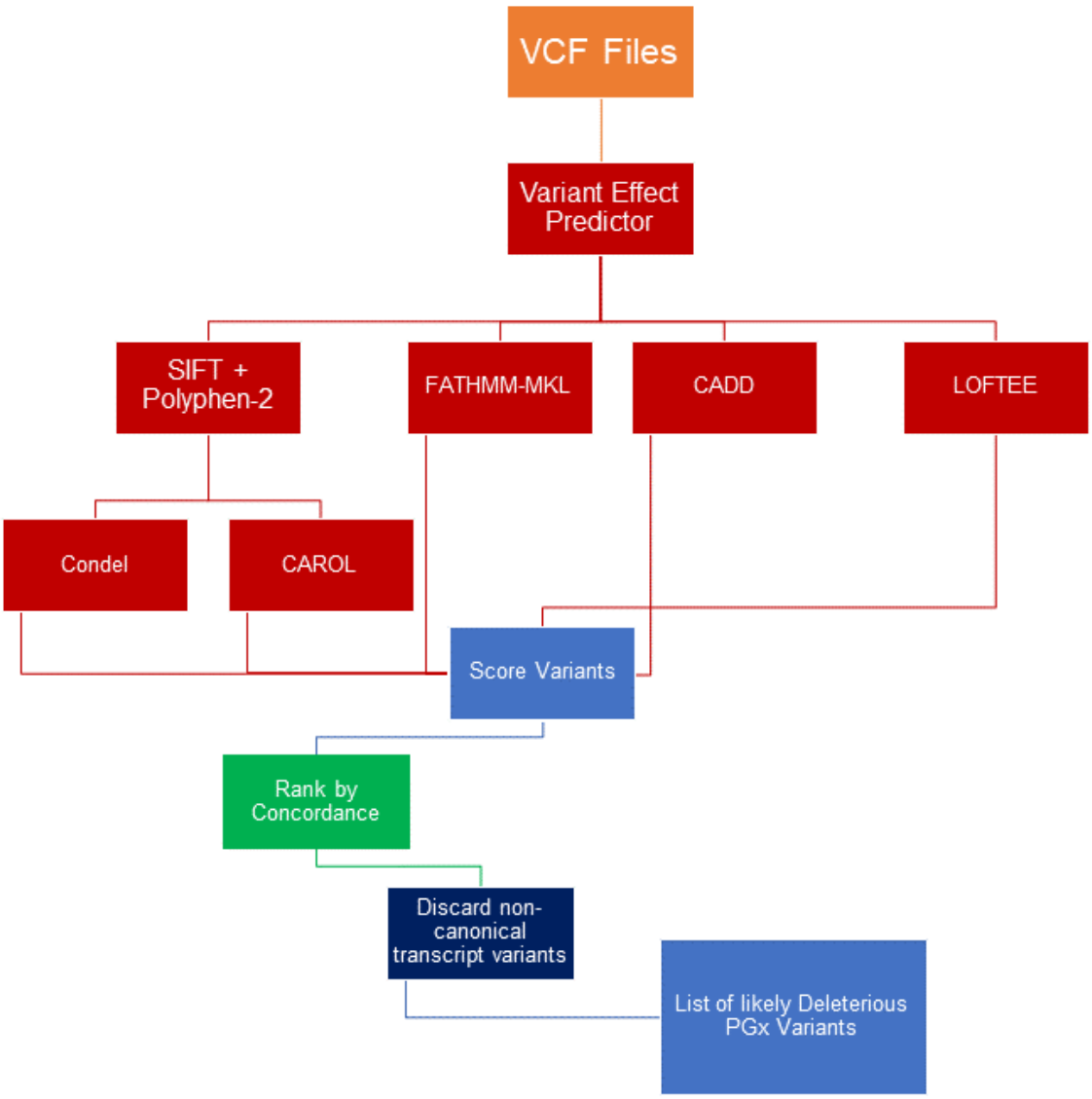


Figure 2. 2 Workflow of VCF files into bioinformatics programs used in this study to select potentially deleterious variants

The workflow used combines several bioinformatics tools aimed at annotation of variant function. The combination of multiple tools allows for more robust prediction and can reduce uncertainty of the effect of variants where tools may not be in concordance. SIFT and Polyphen-2 are the most

widely used tools for variant annotation in coding sequences (they do not provide annotation for non-coding sequences like 5' and 3'UTRs, introns and flanking regulatory sequences). As VEP is a position-based annotation tool, it may annotate variants that are located in overlapping regions (e.g. for two genes in the same genomic position where one may be transcribed from the forward strand and other from the reverse), such that the annotation may apply to the other gene which was not selected for this study. Variants not present in canonical transcripts for the 13 genes of interest were removed from the dataset. Canonical transcripts were defined by Ensembl (Yates *et al.*, 2016) in a hierarchy as follows: 1- those which are the longest CCDS translation of their respective gene with no stop codons; 2 – the longest Ensembl/Havana merged translation with no stop codons; 3- The longest translation with no stop codons; 4 – if no translation product, the longest non-protein coding transcript.

2.2.1 SIFT

SIFT can predict if an amino acid change is likely to impact protein function based on orthologous sequence homology and the chemical similarity between alternate amino acids (Kumar *et al.*, 2009). In a general run, SIFT accepts a user submitted query sequence, then queries protein databases for related sequences, and builds a multiple sequence alignment. SIFT then calculates a conservation value and a probability score for each position in that sequence. These probabilities are then used to assess the likelihood of a deleterious consequence should a substitution occur at a given position. These measures are used to create scaled scores. Scores closer to zero are more likely to induce a deleterious outcome. SIFT has been integrated into VEP via pre-calculated score tables for all proteins in the Uniprot database. When analysing VCF files, VEP will retrieve the SIFT score for all non-synonymous variants listed in the file.

2.2.2 Polyphen-2

Polyphen-2 combines sequence homology, protein family annotations, and 3D protein structures (where available) to predict the outcome of amino acid substitutions (Shihab *et al.*, 2013). Polyphen-2 accepts an amino acid sequence as input, then performs a multiple sequence alignment with highly homologous sequences, and characterises features based on sequence and structural identity via a greedy iterative algorithm. The most important predictive features provide information on how likely two alleles are to occupy a site given the pattern of amino acids in the multiple sequence alignment, how distant the wild type protein is from the alternate deviation, and whether the variant occurs at a hypervariable site. Polyphen-2 is incorporated into VEP, also via

score tables. These tables were produced via running Polyphen-2 on the entire Uniprot non-redundant protein database, along with the DSSP (a secondary structure database from the Protein Data Bank (PDB). SIFT and Polyphen-2 do not always yield the same functional prediction outcome for the same variant. One may predict a variant to be deleterious and the other will predict the same variant to be benign. To assist in the interpretation of such contradictory outcomes, tools such as CAROL (Lopes *et al.*, 2012) and Condel have been developed that can interpret SIFT and Polyphen-2 scores together. These tools automatically retrieve the SIFT and Polyphen-2 results from VEP and calculate a combined score for the variants as the program runs.

2.2.3 CAROL

The Combined Annotation scoring Tool - CAROL uses a weighted Z method to combine the probabilistic scores from the SIFT and Polphen-2 outputs. This method compares the normal standard deviation of scores produced by the tools to a weighted measure of the score for that specific variant. More weight is assigned to scores that indicate a higher probability of a deleterious variant. These weights can then be combined to form a joined predictive effect. CAROL provides higher statistical accuracy than SIFT or Polyphen-2 alone based on overall positive predictions (Lopes *et al.*, 2012).

2.2.4 Condel

Condel makes use of a weighted average scoring model to combine positive prediction scores from SIFT and Polyphen-2. The weights are reflective of how extreme the value of the score is within the distribution produced by a given tool. For a deleterious variant, the weight (the probability that it is not a false positive) increases with the score, penalising variants close to tool cut-offs of deleteriousness. For variants predicted to be neutral, smaller scores produce smaller weights (the probability that it is not a false negative) and inflict a higher penalty. If both tools produce the same score for a variant, more weight will be given to score within the upper distribution of deleteriousness predictions, while less weight will be given to the tool which scores within the ranges of neutral predictions. This method allows the use of weights to combine the underlying probabilities of the tools, producing a consensus prediction.

2.2.5 CADD

The CADD framework is a tool for integration of multiple genome annotations and provides a score for human single nucleotide variants (SNV), and small insertion or deletion events. This tool

provides scores for both coding and non-coding variants. CADD provides a measure of deleteriousness of the variant on the function of the gene it is related to. This framework was built on identified differences between human genomes, and an ancestral genome constructed from human and chimpanzee sequences, where the derived allele frequency in humans was at least 95%. Multiple annotation methods were applied to these sites, as well as sets of simulated variants to serve as controls. The annotation methods used covered a wide range of data sources from sequence databases to regulatory datasets such as ENCODE (Dunham and Kundaje, 2012). The annotations were used to create models which were used to assess and score (named C scores) all possible variants in the human reference genome (GRCh37). To simplify interpretation, the scores were defined into scaled indexes, called phred-like scaled C scores. Scaled C scores are based on the rank of the C score of each variant compared to all other possible SNVs. The top 10% of all C scores (the least likely to be observed in humans based on this model) were assigned a scaled C score of 10 or greater. The top 1% of C scores were assigned scaled scores of 20 or greater, and the top 0.1% were assigned scores of 30 or greater. CADD was included as a VEP plugin, by adding a precomputed dataset of scores for all SNVs in the human reference genome (http://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome _SNVs.tsv.gz)

as well as an indel dataset

(http://krishna.gs.washington.edu/ download/CADD/v1.3/InDels_inclAnno.tsv.gz).

2.2.6 FATHMM-MKL

FATHMM-MKL can predict the function of coding as well as non-coding variants. This is useful for characterising regulatory variants based on their potential effect on gene expression. This tool uses a machine learning approach to annotate variants with data derived from ten datasets which include regulatory information from ENCODE, large sequence conservation datasets, GC content and genome segmentation loci. The tool uses multiple kernel learning classifier to quantify the similarities between data points to assess if a SNV is likely to be functional. FATHMM-MKL can predict the potential effects of non-synonymous variants in similar accuracy to CADD, but out performs it when classifying non-coding sequence variants. This tool was included as a VEP plugin with the addition of a precomputed score dataset for GRCh37 (http://fathmm.biocompute.org.uk/database/fathmm-MKL_Current.tab.gz).

2.2.7 LOFTEE

LOFTEE (https://github.com/konradjk/loftee) is a tool designed to plugin to VEP and annotate Loss of Function (LoF) variants, which completely knockout a gene's functional ability. The tool is designed to identify stop gain, stop loss, frameshift inducing, and splice site disrupting variants. It assigns low confidence predictions based on criteria such as: stop gain or frameshift variants that occur in the last 5% of their respective transcript, or in exons surrounded by non-canonical splice sites. For splice site variants, LOFTEE removes those that are in small introns, those that are in introns surrounded by non–canonical splice sites, and those that are "corrected" by nearby in-frame splice sites. For all LoF variants, LOFTEE will remove those where the purported variant is the ancestral state, as it is unlikely that a LoF variant is the major allele in the ancestral sequence.

Table 2. 1 Summary of functionality of annotation tools used in this study

Tool	Function	Pre-computed dataset	Type of variants annotated	Database type	Scoring	Reference
Variant Effect Predictor	Annotates variants into types*. Serves as base for other programs to input.	No		Ensembl Transcript Database	None.	(McLaren *et al.*, 2016)
Polyphen-2	Scores based on likelihood of harmful amino acid substitution	Yes	Coding	Multiple sequence alignment matrices	Scaled. Variants scored closer to 1 are more likely to be deleterious and closer to 0 more likely to be benign.	(Shihab *et al.*, 2013)
SIFT	Scores based on conservation of amino acids	Yes	Coding	Multiple sequence alignment matrices of highly conserved sequences	Scaled. Variants scored closer to 0 are more likely to be deleterious, while closer to 1 more likely to be benign.	(Kumar *et al.*, 2009)
CAROL	Statistically combines SIFT and Polphen-2 scores	No	Coding	N/A	Combines scores to assist with conflicting SIFT and Polyphen-2 predictions.	(Lopes *et al.*, 2012)
Condel	Calculates a consensus deleteriousness score for SIFT and Polyphen-2 predictions	Yes	Coding	Precomputed SIFT and Polyphen-2 score matrices	Combines scores to assist with conflicting	(González-Pérez and López-Bigas, 2011)

					SIFT and Polyphen-2 predictions.	
FATHMM-MKL	Predicts Functional Consequences of both non-coding and coding sequence variants	Yes	All variants (coding, non-coding, regulatory, etc.)	Precomputed prediction dataset of FATHMM scores for GR37 reference variants	Scaled. Variants scoring above 0.5 are considered deleterious, while those closer to 1 or 0 are of a higher confidence of being deleterious or benign respectively.	(Shihab *et al.*, 2015)
CADD	Combines multiple annotation methods to assign a relative prediction score	Yes	All variants (coding, non-coding, regulatory, etc.)	Precomputed scores for all possible nucleotide substitutions in the GR37 reference dataset.	Scaled. Higher scores indicate greater confidence in prediction.	(Kircher *et al.*, 2014)
LOFTEE	Predicts loss of function variation based on ancestral sequence identity	No	Coding	Human Ancestral Sequence used for comparison.	Binary. Identified variants are classified into classes of High Confidence or Low Confidence only.	https://github.com/konradjk/loftee

VEP classifies variants into defined types, of which the more relevant for this study are: missense, synonymous, 5'UTR, 3'UTR, intronic, start lost, stop lost, stop gain, frameshift, splice acceptor/donor variants and upstream/downstream gene variants. A full list of classifications is available at: https://www.ensembl.org/info/genome/variation/predicted_data.html

These classifications are key to evaluation of LoF variants. Premature stop codons indicated by "stop gains" induce a shortened protein primary sequence (or nonsense mediated decay). Frameshift variants are those that shift the reading frame of transcription machinery after the start site, leading to a disordered sequence of amino acids following the variant position. Variants that occur at splice acceptor or donor sites can lead to an intron being retained into mRNA, which may then express additional amino acids not meant to be in the normal protein structure.

2.2.8 Scoring process

The predictive scores produced via the tools were used to assess the likelihood of a variant having a functional impact on its relevant gene. A more detailed overview of score interpretations for the toolsets is available in Appendix Table 1. The most relevant variant types that were explored in this study are missense and LoF type variants, as these have direct impact on protein function. Missense variants were evaluated with the significance cut-offs listed in Table 2.2. LoF type variants are evaluated in terms of confidence of their effect, for which LOFTEE was used to assess if these were high confidence or low confidence. Non-coding and regulatory variants are more difficult to assess, but as these types are known to have PGx related effects, FATHMM-MKL and CADD scores are useful to compare these with previously reported variants with clinical effects.

Table 2. 2 Score significance cut-offs for prediction of functional impact for missense variants

Toolset used for functional prediction of Missense variants	Score range	Score limit for significant functional impact
SIFT	0-1	≤ 0.09
Polyphen-2	0-1	> 0.6
CAROL	0-1	> 0.981
Condel	0-1	> 0.455
FATHHM-MKL (Coding)	0-1	> 0.7
CADD (phred)	0-99	> 18

The score significance limits are used to distinguish missense variants that are likely to have a small impact on protein structure from those which are predicted to largely alter the overall structure and possible function. These were taken as defined via the toolset authors. To filter missense variants that have conflicting predictions across tools, consensus was evaluated to distinguish variants that are of high confidence to have a functional impact (Table 2.3). This ranking system cannot be applied to LoF and non-coding variants. They can be assessed via estimations by LOFTEE, and via CADD and FATHMM-MKL respectively.

Table 2. 3 Description of the ranking system used to assess variants in terms of their likelihood of functional impact.

Variant Class	**Ranking System**		**Variants selected for subsequent analysis**
Missense	**All six tools in concordance with significant prediction scores**	**Four or Five tools with significant predictive scores**	
High Confidence Functional Impact	Yes	No	Yes
Likely Functional Impact	No	Yes	Yes
Low Confidence of Functional Impact	No	No	No
Loss of Function	**LOFTEE High Confidence**	**LOFTEE Low Confidence**	
High Confidence of Effect	Yes	No	Yes
Low Confidence of Effect	No	Yes	No
Non-Coding	**CADD (score > 18)**	**FATHMM-MKL (score > 0.7)**	
Likely Functional Impact	Yes	Yes	Yes*
Low confidence of Functional Impact			No

*Due to only two tools to assess their impact, non-coding variants that are likely functional were compared via a list of known variants from PharmGKB (see section 2.2.9) and would only move to subsequent analyses if they have been previously reported for functional effect.

This method defines missense variants with complete consensus of prediction into "High Confidence" of functional impact, while variants with above four toolsets predicted are only of "Likely Functional" impact. Missense variants with fewer than four toolsets in consensus, and LoF variants with low confidence estimations from LOFTEE were filtered (removed) from the datasets prior to frequency analysis.

2.2.9 PharmGKB database mining for known variants

To identify variants previously determined to be pharmacogenomically relevant and that would have been filtered out using the approach above, the PharmGKB database was used to identify variants in the 13 genes from this study. PharmGKB provides a wealth of data available for pharmacogenomics use. There are multiple datasets available use, containing lists of genes, drugs and phenotypes, but also a list of PharmGKB annotated variants that are also tracked in dbSNP (have been previously identified). The dataset for all variants was obtained from https://www.pharmgkb.org/downloads, (accessed 1/11/2017) and contains the variant ID, its gene ID and symbol, and the count of "Variant Annotations" and "Clinical Annotations" for that variant. Variant Annotations describe allele-based associations, and Clinical Annotations describe genotype-based associations. Clinical annotations recorded in PharmGKB are scaled in four levels, which are based on the strength of evidence for variant association. These levels are distinguished via the following criteria: Level 4 annotations are those based on preliminary evidence from single cases or *in vitro* studies; Level 3 annotations are supported by statistically significant studies which are not yet replicated, Level 2 annotations have been replicated but some studies may not show complete statistical significance, as well as those reported with strong effect sizes in known pharmacogenes; Level 1 annotations are those which have been endorsed by medical society guidelines for drug use, such as those issued by the CPIC.

Variants for the PGx genes assessed in this study were extracted in a manner to include only variants that have five or more "Variant Annotations", or those with one "Clinical Annotation" as well as at least three Variant Annotations. This ensures that significant counts of sources or records are present to sufficiently characterise the variant and forms a refined list which was used for comparison with datasets assessed in this study. VEP provides an annotation for existing variants (rsID), and these were used for a match comparison to compare the population datasets assessed in this study with the variant list from PharmGKB. Any variants which overlap between these sets are of relevance to discuss in terms of PGx impact in Africa as these have prior evidence of their effect outside of only bioinformatics-based prediction methods.

2.3 Allele frequency determination and Statistical Comparisons

PLINK v1.9 (Chang *et al.*, 2015) was used to call allele frequencies of variants and their counts within the population datasets chosen for analysis. This tool accepts the VCF format as input and is suitable for analysis of large datasets. An example script of running PLINK is shown in the

Appendix Script 7.2.2. Multiallelic variant frequencies will only be called for the most common allele in these cases. Allele frequency data from the KGP global population groups, European (EUR), Admixed American (AMR), East Asian (EAS), and South Asian (SAS) were retrieved in VEP (see Appendix Script 7.2.1) and used for comparison with the grouping of continental Africans, which was formed by averaging the allele counts per variant for the five KGP African populations and the three AGVP African populations.

To compare if allele frequencies were significantly different, the Chi-square test and the Fisher's exact test were used. The Fisher's exact test was used for rare variants not meeting the criteria to accurately compare using a Chi-Square test, and to compare populations in a pairwise manner. However, the Fisher's exact test is computationally expensive when comparing multiple populations together for common variants – which occurs when assessing overall difference between the eight African populations. For these cases, the Chi-square test was used when the allele count was at least five. The statistical calculations were performed in R v3.2.2 (Fire Safety) (R Core Team, 2015), with the "stats" package installed (Appendix Scripts 7.2.5-7.2.6).

Continental African population frequencies were compared in an overall test, to assess if there is a significant difference between the eight populations together. These populations were also compared in a pairwise manner, comparing each pair of populations to assess intra-African differences. Significance was defined as $p<0.05$ for pairwise population comparisons, and $p<0.006$ (0.05/8) for the overall test for difference between the eight continental African populations to correct for multiple testing.

The frequencies for KGP global populations were compared with continental African population grouped frequencies in a pairwise manner, to establish which of these populations are significantly different from Africans. An overall test for these five inter-continental populations together is unnecessary as this study is focused on difference to Africans, not between the other global populations. Significance was defined as $p<0.05$ for these inter-continental population comparisons.

Chapter 3: Results

This study aimed to identify genes involved in PGx effects for the main cancer types affecting Africa and annotate variation in these genes by using publicly available WGS datasets for African populations. The annotation provided more information about the variant, including predictive scores for their likelihood of having a functional impact on their respective genes. These annotations were used to create a list of variants of high likelihood to have a pharmacogenomic related effect, which was compared to a list of known variants obtained from PharmGKB.

To select genes to evaluate, PharmGKB records were searched via cancer type to return genes that have previously reported PGx related effects to drugs the drugs used in the treatment of those types (Table 3.1). From these, a set of 13 genes was selected for analysis, based on the data reported in Table 3.1 and Appendix Table 2, which aimed to overlap cancer types and the maximum number of drug types. Genes selected include the major PGx groupings: drug metabolisers, drug transporters, and those that form high risk genotypes. Genes encoding drug metabolising enzymes are: *CYP19A1, CYP1B1*, *CYP2D6, CYP3A4, CYP3A5, DPYD,* and *GSTP1*. Genes encoding key drug transporters are *ABCB1* and *SLC19A1*. Genes forming high risk genotypes can be arranged in two categories, specific drug targets: *ESR1* and *CYP19A1*, and DNA repair enzymes: *XRCC1* and *XRCC5.*

Five genes, *CYP2D6, CYP3A4, CYP3A5, DPYD* and *ESR1*, although only reported in breast cancer related PGx, were included due to special characteristics. Breast cancer rates and mortality are rising substantially, and breast cancers are known to be more severe in African women. *CYP2D6* and *ESR1* are key genes to the function of tamoxifen, a widely used drug to prevent breast cancer recurrence. *CYP3A4* and *CYP3A5* metabolise a large number of drug types (see Section 1.5.2). Adverse reactions from 5-FU, a drug which must be metabolised by DPYD are severe and are often lethal to the patient.

Table 3. 1 PGx related genes according to cancer type obtained from PharmGKB records for neoplasm types (those highlighted in yellow were investigated further)

Uterine Cervical	Prostatic		Lung	
GSTM1	*ABCC6*	*SULT1A1*	*ABCG2*	*HOTTIP*
GSTP1	*AKR1C3*	*SULT1A2*	*AHR*	*HYKK*
GSTT1	*ATP7A*	*SULT1C4*	*AKT1*	*LRP1*
SLC19A1	*CHST3*	*TANC1*	*AQP2*	*MALAT1*
XRCC1	***CYP19A1***	*VEGFA*	*AQP9*	*MEG3*
	CYP1B1		*ATP7B*	*NR1I2*
	CYP2C8		*CDKN2B-AS1*	*PON1*
	CYP4B1		*CHRNA3*	*POU2F2*
	NAT2		*CYP2A6*	*PTEN*
	ORM2		*EIF3A*	***SLC19A1***
	PPARD		*GGH*	*SLC2A1*
	PTGS2		*H19*	*SLCO1B3*
	RPL13		*HLA-DOB*	*SULT1A1*
	SLC10A2		*HMGB1*	*SULT1A2*
	SNORD68		*HMGB2*	*TMEM205*
	SPG7		*HOTAIR*	***XRCC5***

Oesophageal	Breast					
ABCB1	***ABCB1***	*CYP2A6*	*EPHA5*	*IGF1*	*SLC28A1*	*WRAP53*
ERCC1	*ABCC1*	*CYP2B6*	*EPHA6*	*IGF1R*	*SLC28A3*	*WWOX*
ERCC2	*ABCC2*	*CYP2C19*	*EPHA8*	*IGFBP3*	*SLC29A1*	*ZNF423*
GNAS	*ABCC4*	*CYP2C8*	*EPM2AIP1*	*IL17F*	*SLC37A1*	
GSTP1	*ABCC5*	*CYP2C9*	*ERBB2*	*IRS1*	*SLCO1B1*	
TYMS	*ABCG1*	***CYP2D6***	*ERCC1*	*MAP4K4*	*SOD2*	
XRCC1	*AKR1C3*	***CYP3A4***	*ERCC2*	*MDM4*	*SPIDR*	
Myelomas & Lymphoma	*ALDH1A1*	***CYP3A5***	***ESR1***	*MISP*	*SULT1A1*	
ABCB1	*C10orf11*	*DAPK1*	*ESR2*	*MTHFR*	*SULT1A2*	
CD3EAP	*CBR1*	*DCK*	*FCGR2A*	*NCF4*	*TCL1A*	
COL1A1	*CBR3*	*DGKI*	*FCGR3A*	*NOS3*	*TNFRSF11A*	
CTLA4	*CCND1*	*DHFR*	*FGD4*	*NQO1*	*TNFRSF11B*	
CTNNB1	*CD3EAP*	*DLGAP1*	*FGFR2*	*NQO2*	*TNFSF11*	
CYP2C8	*CDA*	*DNAH12*	*FGFR4*	*PDE4B*	*TOP2B*	
ERCC1	*CES2*	*DOCK4*	*FLT1*	*PRKG1*	*TP53*	
GSTT1	*CLCN6*	***DPYD***	*FLT3*	*RAC2*	*TUBB1*	
MMP2	*CMPK1*	*E2F7*	*FZD3*	*RBFOX1*	*TUBB2A*	
PPARG	***CYP19A1***	*ECT2L*	*GLDN*	*RRAS2*	***TYMS***	
PPP1R13L	*CYP1A1*	*EGLN3*	***GSTP1***	*RRM1*	*UGT2B15*	
SLC7A5	*CYP1A2*	*ENOSF1*	*HLA-DQA1*	*SETD4*	*UGT2B7*	
SPP1	***CYP1B1***	*EPAS1*	*HLA-DRB1*	*SLC22A16*	*VEGFA*	
TNFRSF11A						
TNFRSF11B						
XRCC5						
ABCC2						
CYP2C19						
UGT1A1						

3.1 Variant Characterisation

3.1.1 Gene based variant characteristics

Variants within and around the selected 13 genes (Table 3.2) were extracted from the complete variant call file for each individual in each population. Once annotated, variants from all populations were filtered to generate a set of unique non-redundant variants by removing all duplicate variants (from non-canonical transcripts). The genic location of the set of non-redundant variants and the frequency of variants in each genic location are shown (Fig 3.1 - 3.3). This is useful to assess the broad spread of gene-associated variants analysed for each respective gene region per gene. The genes were of different lengths, as were the relative lengths of the different gene regions. Essentially, the larger genes had more variants than the smaller genes.

Table 3. 2 Gene length defined for the canonical gene transcript (start to end of transcription)

Gene Name	Canonical Transcript Length (bp)
ABCB1	209,389
CYP19A1	130,553
CYP1B1	8,671
CYP2D6	4,407
CYP3A4	26,589
CYP3A5	31,802
DPYD	843,280
ESR1	295,720
GSTP1	3,065
SLC19A1	28,695
TYMS	15,974
XRCC1	32,966
XRCC5	98,839

Only the gene length for the canonical transcript for each gene as provided by Ensembl is shown. *DPYD* and *ESR1* are the largest genes assessed. The genes for most cytochrome P450 family of drug metabolising enzymes are relatively small, with the exception of *CYP19A1*, which is a drug

target in this assessment. The counts of variants in Figures 3.1-3.3 below were done by summing up the totals of all variants detected in the eight African populations and presenting them as the total of unique (non-redundant) variant for each gene. The figures are useful to characterise variant types in each gene and to assess how these may influence the number of variants predicted to have a likely functional impact.

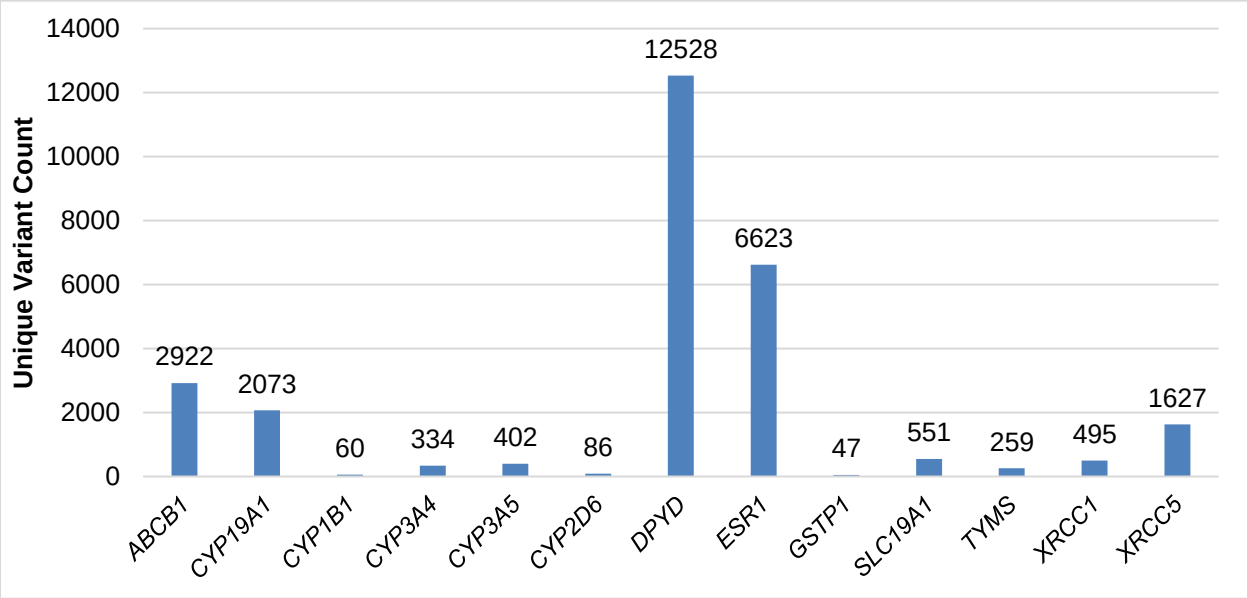


Figure 3. 1 Total count of intron variants per gene summed across all populations analysed

The majority of variants were intronic variants. An absolute count of variants per variant categories is shown elsewhere (Appendix Table 3). Proportions (intronic/total) vary greatly between genes, with: <95% for *ABCB1, CYP19A1, DPYD, ESR1,* and *XRCC5*; 64% - 87% for *SLC19A1, TYMS, XRCC1, CYP3A4* and *CYP3A5,* and 43%, 46% and 24% for *GSTP1, CYP2D6* and *CYP1B1* respectively. It appears the larger the gene, the higher the relative proportion of intronic variants.

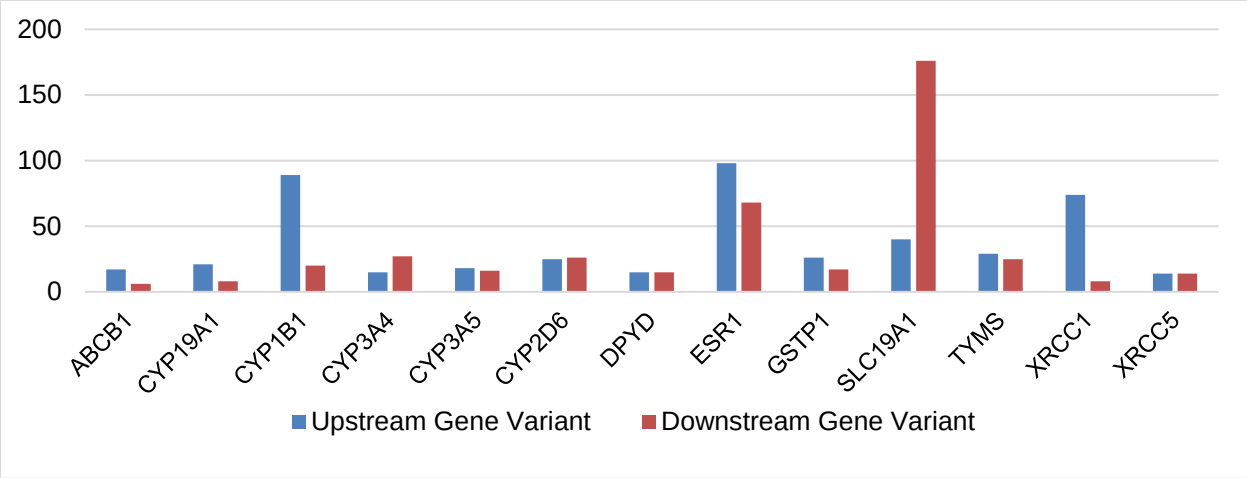


Figure 3. 2 Variant counts for upstream and downstream flanking regions (1000bp upstream and 1000bp downstream of the regions from the start and end of transcription, respectively)

Figure 3.2 shows the variants for a fixed number of nucleotides flanking the transcribed regions. Since their length is the same for each gene, the number of variants for each gene is not relative to the size of the gene. Upstream variants are in a region enriched for regulatory variants potentially affecting gene transcription. They may or may not have a functional impact. Interestingly, three genes appear to have significantly more upstream gene variants (*CYP1B1*, *ESR1* and *XRCC1*) compared to the majority of genes. *SLC19A1* displays the highest count of downstream gene variants with *ESR1* the second highest and the rest having a relatively small number of downstream variants. *ESR1* is the only gene that has a high number of variants both upstream and downstream.

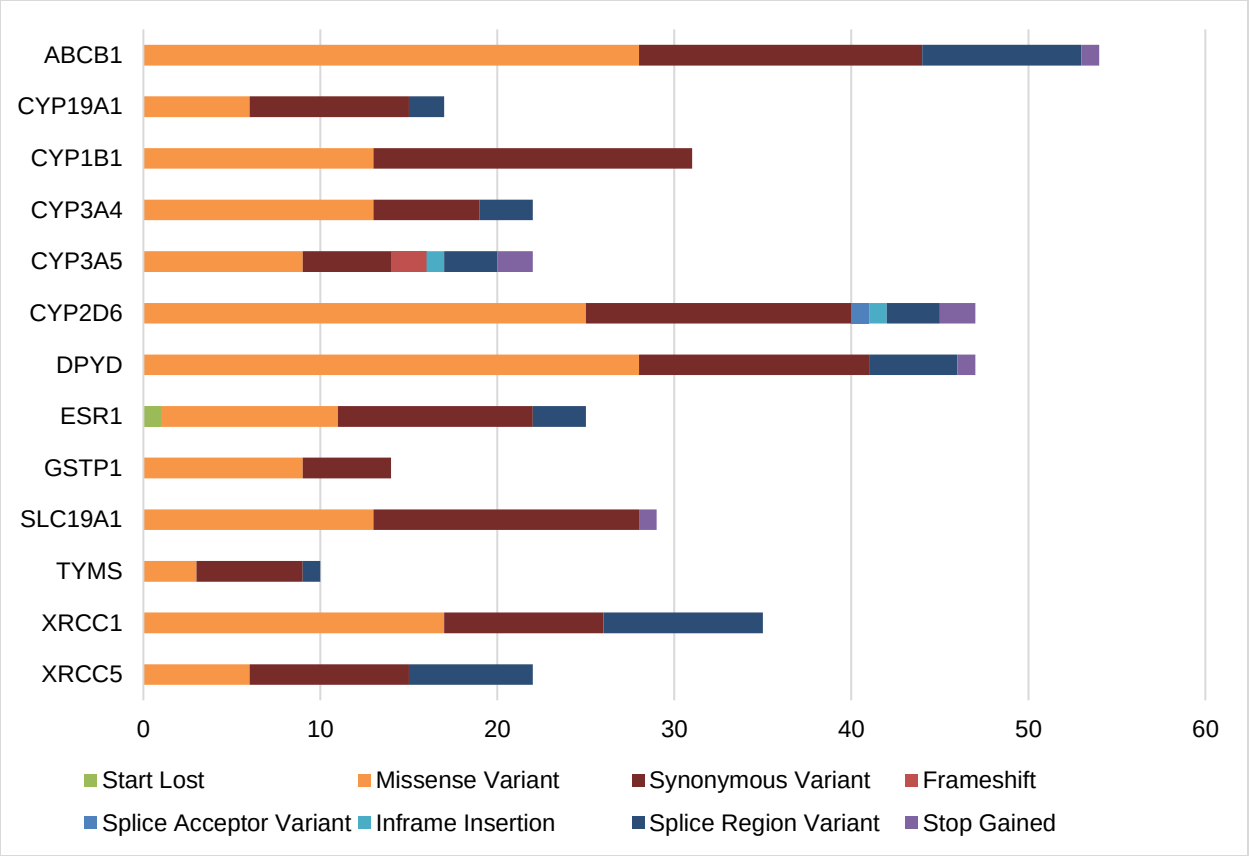


Figure 3. 3 Variant counts for coding region variants

Coding variants (with the exception of synonymous variants, that very rarely are pathogenic) are most likely to induce an effect on the gene product, as they may directly impact the outcome of a protein sequence. LoF variants, as implied by the name, are the most certain to cause effects and they include stop gain and losses, and splice variants. Stop gain mutations can result in nonsense-mediated decay and therefore no protein at all. Other variants cause modification of the primary amino acid sequence, such as missense variants, or in frame insertions or deletions. Splice region variants occur at the exon/intron boundaries usually within 1-3 bases of the boundary in either the exon or intron sequence. Scores for the predictive toolsets have been analysed in more detail for these variant types, and the highest-ranking variants of these are reported in the Section 3.2. Figure 3.3 shows the number of coding region variants for each gene. The most common variants are missense and synonymous for most of the genes and in a minority of genes (e.g. *ABCB1*, *DPYD*, *XRCC1* and *XRCC5*) there are a large number of splice region variants. *ESR1* has the only start loss variant and several genes have one or two stop gain or nonsense variants. Generally, as expected, the large genes have more coding variants.

3.1.2 Population based variant characteristics

In this section the variants are described in the context of each of the African populations. It examines the different types of variants per gene per population by counting the variants as either present or absent in a population. In this section the frequency of the variant in a population is not taken into consideration. The frequency data (assessed in later sections) are only provided for the variants of interest in terms of function.

The total variant count by individual population is displayed below in five figures, each examining a different class of variant. These totals represent the count of unique variants within the original VCF files, within each respective gene region per population, which would go on to be annotated for functional prediction (Table 3.4). The count of variants shared between populations are not investigated at this stage of the assessment, but only once annotation has revealed variants of interest. The absolute total of variants per gene (Figure 3.4) are distributed consistently across population, which is a good indicator that VCF files were prepared correctly as no population is under- or over-represented in any gene region. Although the total count of variants per gene are consistent, Figures 3.5-3.8, indicate that although the total count of variants per population may be similar per gene, these totals may be the result of different kinds of variants for each respective population. By comparing Figure 3.1 and Figure 3.5, it is clear that the greatest proportion of overlap between populations is in intronic variants. Other variant types (Figures 3.6-3.8) are not as linearly shared between populations indicating a high degree of variability at a gene level. Rare variants seldomly overlap between population groups (Table 3.3 and 3.5). Counts of other variant types split by population are available in Appendix Table 3.

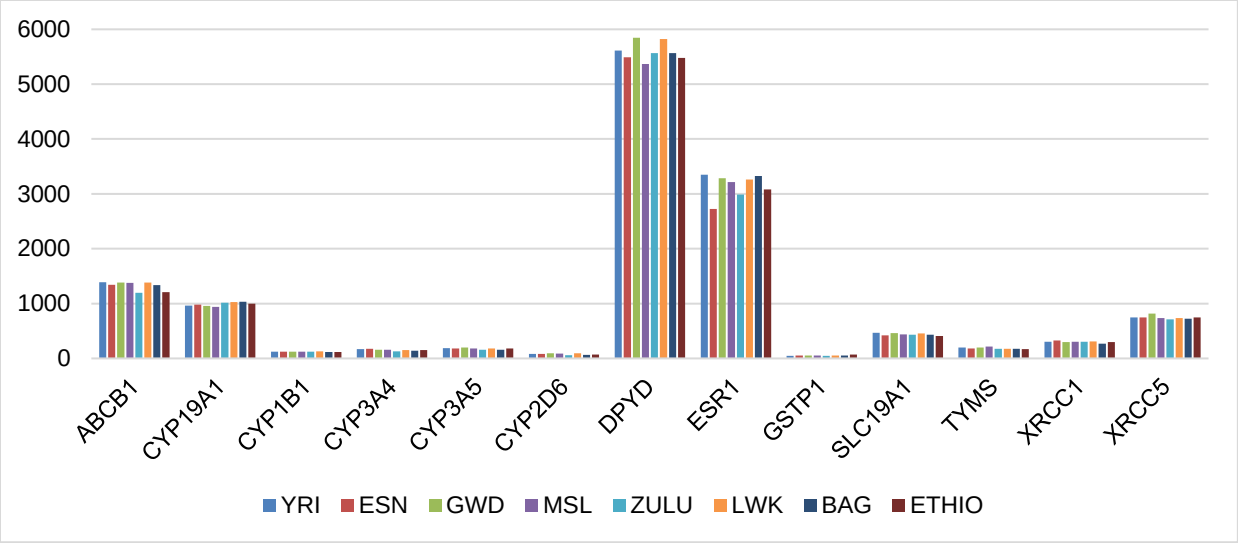


Figure 3. 4 Total count of variants per gene for each continental African population group

These totals apply to unique levels of variants per population group split by gene. This shows that variation across the populations is well represented across the same genes, though there are some differences in specific populations for *DYPD* and *ESR1*. These totals are very large, although it is noted that the greatest contribution to the totals are the intronic variants (Figure 3.1). By comparing this number to Figure 3.1, it is clear that there are substantial levels of overlapping variants between the populations. With *DPYD* as an example, this gene has 12000 intronic unique variants in total, which was formed from ~5300-6000 variants each contributed from eight populations.

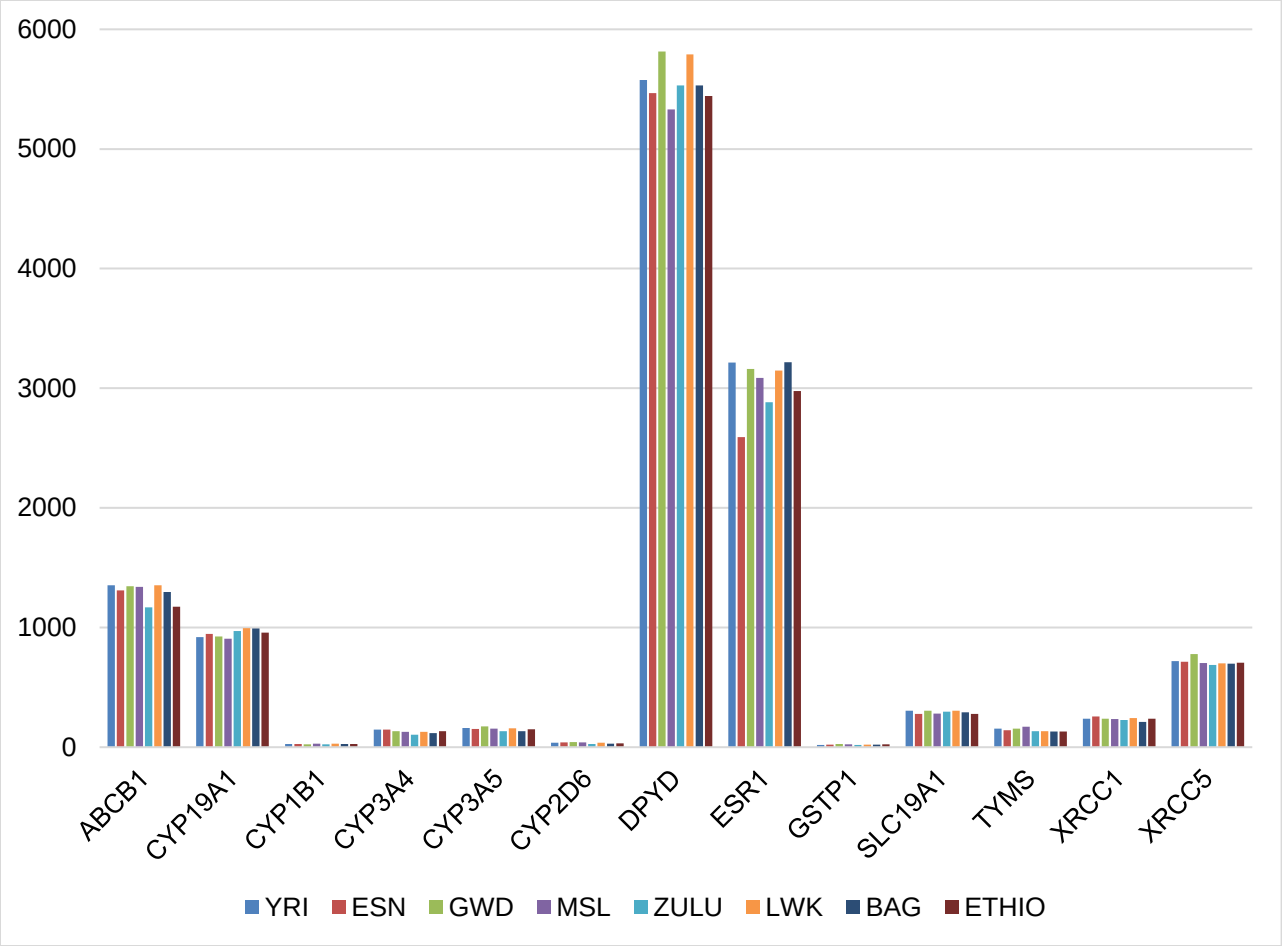


Figure 3. 5 Intron variant count per gene for each continental African population

These variants (Figure 3.5) are found in the intronic regions of the gene. The count of intron variants in these genes forms the main contribution to the population-based differences of total gene count seen in Figure 3.4 and contribute to total count more than any other variant type. The *DPYD* and *ESR1* genes have the greatest amount of intronic variation and are also the largest genes assessed in this study.

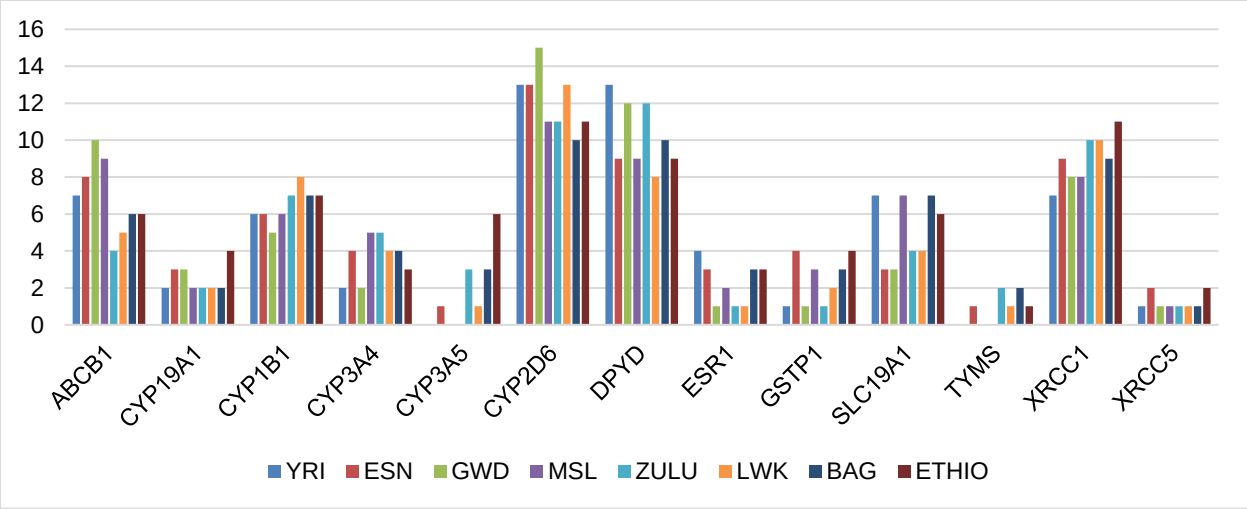


Figure 3. 6 Missense variant count per gene for each continental African population

Missense type variation shows more variation across populations by count, although not in all genes. The *CYP3A5* and *TYMS* genes have sparse missense variation across population groups. Missense variation appears not to be affected relative to gene size. *CYP2D6*, one of the smallest genes has comparable numbers of missense variants to *DPYD*, the largest gene assessed in this study.

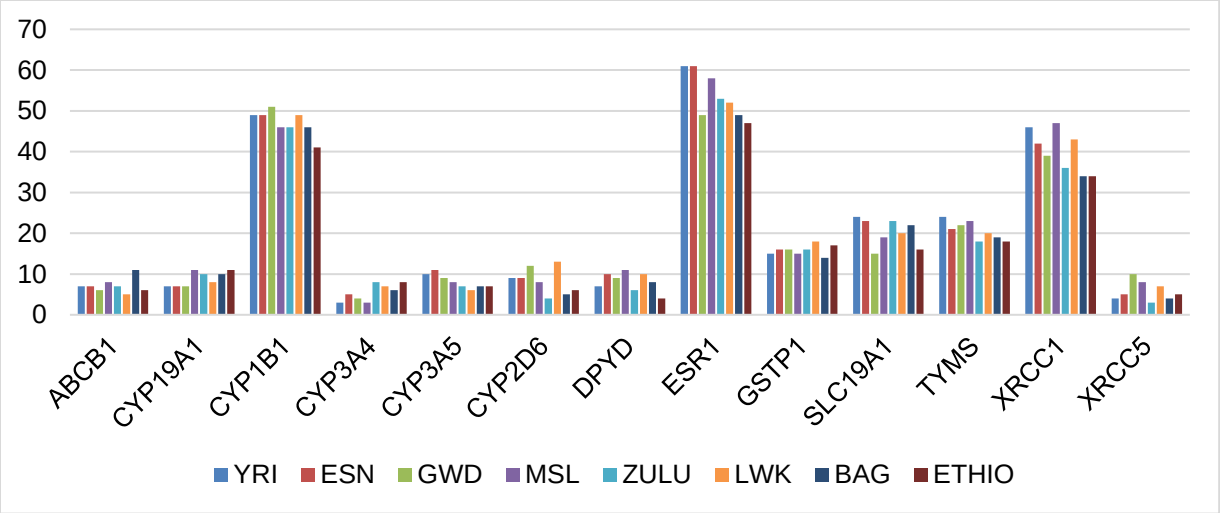


Figure 3. 7 Upstream Variant Count by population

Upstream gene variants are those within the 1000bp flank of the gene leading up to the transcription start site. Upstream gene variants are very consistently distributed across populations

in all gene types, though some genes, such as *CYP1B1*, *ESR1*, and *XRCC1* have two to three-fold more of these variant types than other genes.

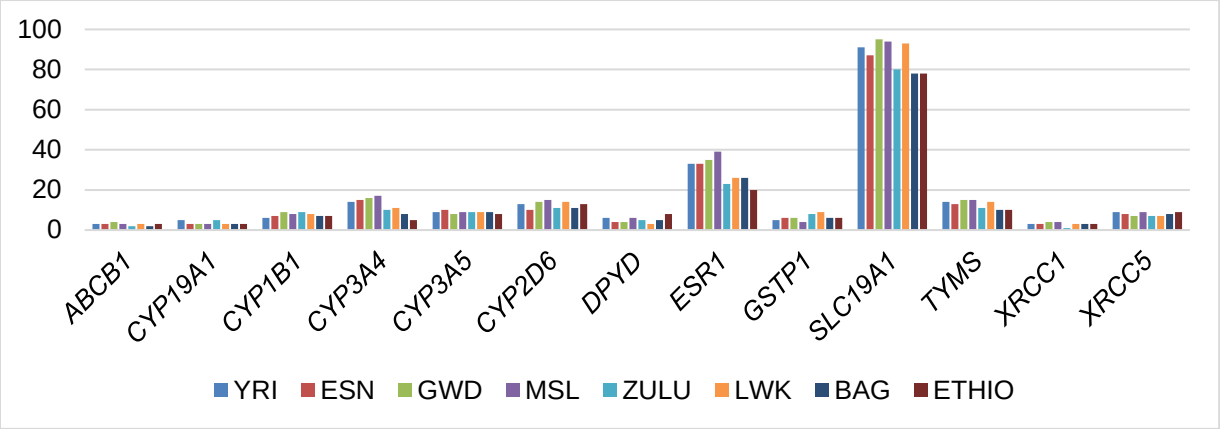


Figure 3. 8 Downstream Variant Count by population

Downstream gene variants are those which lie in the 1000bp flanking region following the transcription end site. The numbers of downstream gene variants are consistent in most genes and populations, with the exception of *ESR1*, which appears to have greater counts in West Africans. *SLC19A1* has the largest count of downstream variants, which tend overlap between populations, as an average of ~90 variants in each population lead to a total of 176 unique variants of this type assessed (Figure 3.2).

3.2 Variants with high likelihood of functional impact

This section describes variants with the highest likelihood of being deleterious to the function of their respective gene products based on prediction scores of annotation tools. When variants are addressed in sections 3.2 through 3.4 in text, they are referred to via their rsID nomenclature which is tied to their effect allele as assessed via functional prediction. These effect alleles are noted in tables and figures and compliment the in text descriptions of variant characteristics. By their nature, LoF variants are highly likely to produce deleterious effects on gene function when they are present. Missense variants can have ranging effects dependent on their location and the amino acid change induced and were classed as likely functional should they meet significant scores for the prediction tools in most toolsets. Missense variants that meet significance score thresholds for all tools used in this study are deemed "high confidence" for functional impact. The effects of variants in non-coding and regulatory regions are more complex, and only two toolsets were used to assess these variants. Additional criteria are needed to be confident of their effect. When frequencies are displayed in the figures to follow, the populations are arranged with West Africans on the left, East Africans on the right, with South African Zulu placed between them. This helps to highlight any regional differences in frequency.

3.2.1 Non-coding and regulatory region variants of likely functional impact

Non-coding variants were assessed via their CADD and FATHMM-MKL scores, along with a comparison to known variants from the PharmGKB database to assess if any of these have prior clinical reports of functional impact. Of the total number, 191 variants were selected based on the scoring criteria defined above, the details are available through Appendix 7.3. The bulk of these variants are intronic, of which the majority lie within the *DPYD* and *ESR1* genes. Potentially functional downstream gene variants were only found for the *SLC19A1* gene, with the other variant types being distributed across other genes.

None of the 191 variants was present in the PharmGKB database, either as a Variant or as a Clinical Annotation, and could therefore not be assessed using this database. Due to uncertainty of the impact of these variants, they were not included further in this study.

3.2.2 Coding Region variants of likely functional impact

Of variants within gene coding regions, 10 LoF variants and 61 missense variants passed the assessment criteria cut-offs for potential functional impact. The numbers of variants per category are shown per gene in Figure 3.9. High confidence missense variants are those for which all tools are in concordance with regard to the predicted outcome on gene function. The scores for both high and lower confidence missense variants are outlined in Table 3.4.

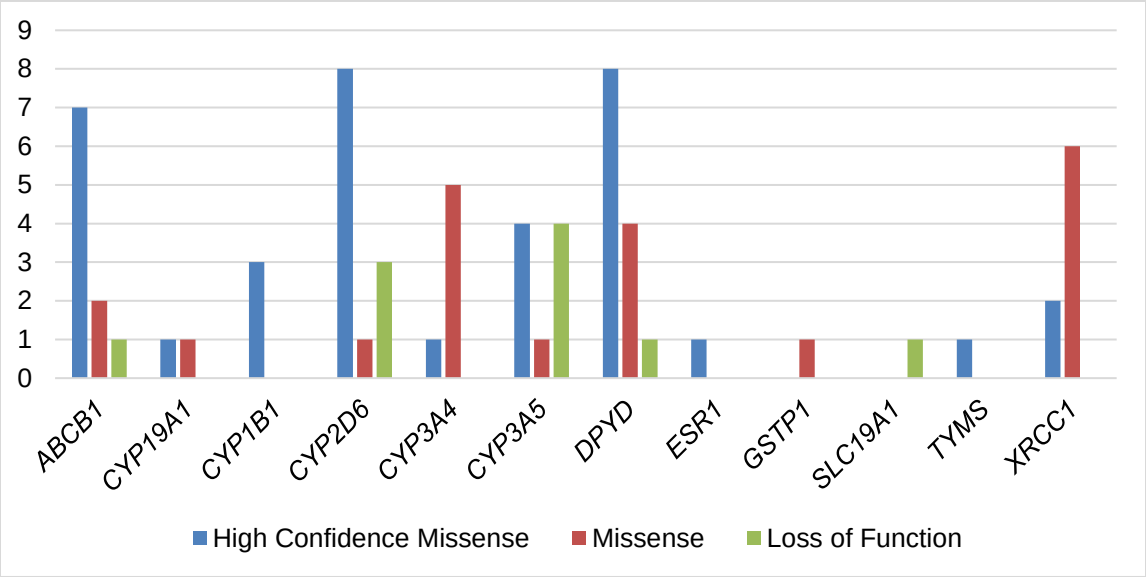


Figure 3. 9 LoF and potentially deleterious missense variant distribution across genes

("High Confidence Missense" variants defined by concordance among all prediction tools; "Missense" variants define by incomplete concordance across tools but a suggestion of deleteriousness)

These variants are not evenly distributed across genes, with the majority occurring in *ABCB1 CYP2D6, CYP3A5* and *DPYD*. The *CYP3A5* and the *XRCC1* genes have high counts of likely missense variants, but relatively few high confidence missense variants compared with *ABCB1 CYP2D6*, and *DPYD*; with no LoF variants. The *XRCC5* gene had no variants surviving significance cut-offs and is not shown in the figure. LoF variants are less commonly found across these genes, with most present in *CYP2D6* and *CYP3A5*.

Table 3. 3 Loss of function variants and frequencies in African populations

Gene	Id	LoF Allele	Consequence	YRI	ESN	GWD	MSL	ZUL	LWK	BAG	ETH
ABCB1	7:87195405	T	Stop Gained	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.004 (1)
DPYD	rs72549310	A	Stop Gained	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP2D6	rs147960066	A	Stop Gained	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
CYP2D6	rs3892097	T	Splice Acceptor Variant	0.055 (12)	0.075 (15)	0.039 (9)	0.041 (7)	0.035 (7)	0.025 (5)	0.03 (6)	0.041 (10)
CYP2D6	rs536109057	A	Stop Gained	0 (0)	0 (0)	0.008 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A5	rs149664815	A	Stop Gained	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (2)	0 (0)	0.008 (2)
CYP3A5	rs41303343	A	Frameshift Variant	0.12 (26)	0.085 (17)	0.137 (31)	0.135 (23)	0.105 (21)	0.116 (23)	0.095 (19)	0.008 (2)
CYP3A5	rs111371159	T	Stop Gained	0 (0)	0 (0)	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A5	rs200579169	C	Frameshift Variant	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.004 (1)
SLC19A1	21:46951785	T	Stop Gained	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	(0)	0.004 (1)

*Number of alleles shown in parentheses

Only one variant (*ESR1* - chr6:152129048, start lost) of the 11 LoF variants originally identified was excluded from this list, due to a low confidence score from LOFTEE. This variant was previously identified in somatic tissue from lung tumour, only in one sample (http://cancer.sanger.ac.uk/cosmic/mutation/overview?id=325271). It therefore has a low-confidence likelihood of producing a loss of function effect of significance. This variant is very rare (allele count of one), present only in the Ethiopian population. The variants in Table 3.3 were all called as High Confidence by the LOFTEE tool and have the greatest likelihood of producing the expected loss of function effect on their respective gene. Of the 10 LoF variants, six were observed only once and one was observed twice. Since the discovery WGS data was low coverage, and not ideally suited for calling low frequency variants, the presence of these variants will need to be confirmed using alternate sequencing technologies. Two variants, rs3892097 and rs41303343, are common to Africans and may be of high relevance to PGx impact. These are represented in the Figures 3.10, and 3.11 to better compare them within Africa and between African and global populations respectively. As noted above, populations have been arranged in order of West, South, and East Africans for comparison.

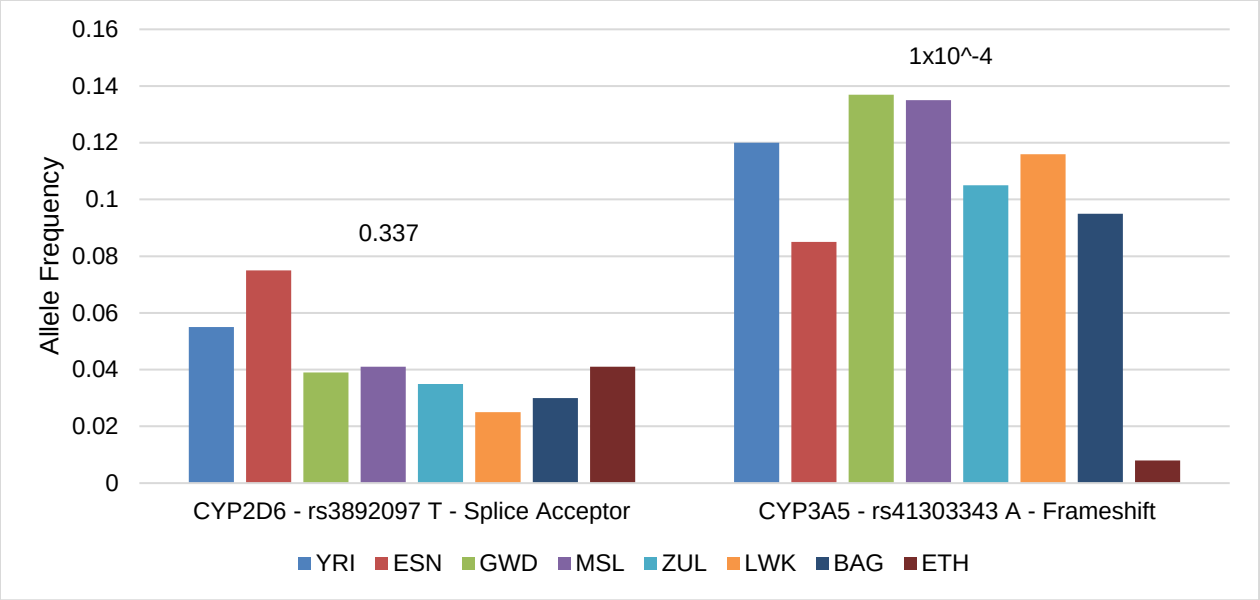


Figure 3. 10 Inter-African frequency variation for the two common LoF variants
The numbers indicated above the bars indicate p-values for differences in alleles for the grouped African populations (described in Methodology 2.3) and pairwise comparison p-values are available through Appendix 7.3.2.

The *CYP2D6* rs3892097 splice acceptor variant is present in all African populations (ranging in frequency from ~0.02 to ~0.07), with some differences in frequencies between the populations (though not different overall at p=0.337) (e.g. the ESN have roughly double the frequency found in most of the other populations). This variant has a high number of clinical reports in the PharmGKB database (Table 3.6). The *CYP3A5* rs41303343 frameshift variant is more common (frequencies ranging between ~0.01 to ~0.14), and also shows differences across African populations. It is notably rarer in the Ethiopians (ETH). It has very few annotations present in the PharmGKB database, all for non-cancer drug types.

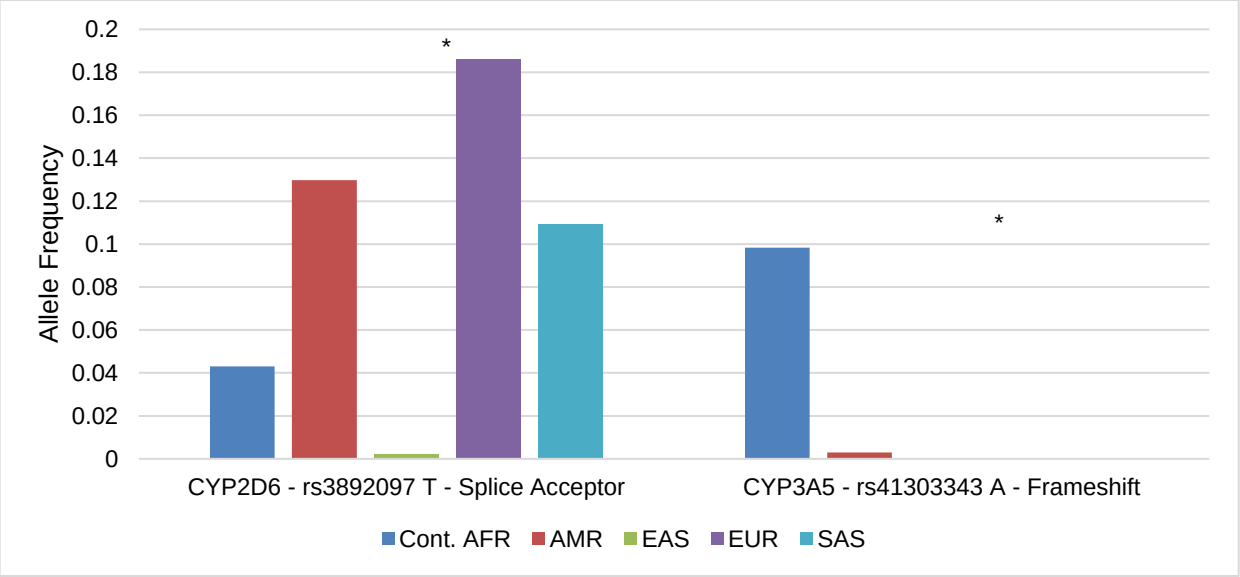


Figure 3. 11 Global comparison of continental African average frequency with KGP continental populations (Admixed American (AMR); East Asian (EAS); European (EUR) South Asian (SAS))

The rs3892097 variant is significantly less common in continental Africans compared to Europeans, Admixed Americans, and South Asians. This variant is, however, almost absent in East Asians. The rs41303343 variant appears to be an African-specific common variant as it is absent from other populations with the exception of the Admixed American population where it occurs at extremely low frequency (possibly due to African admixture).

Table 3. 4 Scores for missense variants with high likelihood of functional impact

GENE	ID	CODONS	AMINO ACIDS	SIFT	POLYPHEN-2	CAROL	CONDEL	CADD PHRED	FATHMM MKL CODING
ABCB1	rs137996914	Cgc/Tgc	R/C	Deleterious (0)	Possibly Damaging (0.677)	Deleterious (0.999)	Deleterious (0.685)	34	0.84
ABCB1	rs142600685	Cgc/Tgc	R/C	Deleterious (0.02)	Possibly Damaging (0.775)	Deleterious (0.986)	Deleterious (0.671)	32	0.92384
ABCB1	rs147487745	Gct/Act	A/T	Tolerated (0.14)	Possibly Damaging (0.674)	Neutral (0.877)	Deleterious (0.489)	26.9	0.90776
ABCB1	rs148718120	tAc/tCc	Y/S	Deleterious (0)	Possibly Damaging (0.66)	Deleterious (0.999)	Deleterious (0.677)	25.8	0.9931
ABCB1	rs200754866	Gcc/Acc	A/T	Deleterious (0)	Probably Damaging (0.938)	Deleterious (1.000)	Deleterious (0.828)	35	0.98608
ABCB1	rs201159898	Acc/Tcc	T/S	Deleterious (0)	Probably Damaging (0.991)	Deleterious (1.000)	Deleterious (0.889)	25.8	0.98384
ABCB1	rs35730308	Tgg/Cgg	W/R	Deleterious (0)	Possibly Damaging (0.761)	Deleterious (0.999)	Deleterious (0.726)	28.4	0.99089
ABCB1	rs535338471	aAt/aGt	N/S	Deleterious (0)	Possibly Damaging (0.784)	Deleterious (0.999)	Deleterious (0.738)	24.4	0.98129
ABCB1	rs572038993	Gct/Cct	A/P	Deleterious (0)	Possibly Damaging (0.7)	Deleterious (0.999)	Deleterious (0.696)	32	0.98978
ABCB1	rs930790323	ttG/ttC	L/F	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	24.9	0.14795
CYP19A1	rs142652579	gGc/gAc	G/D	Deleterious (0.03)	Probably Damaging (0.974)	Deleterious (0.997)	Deleterious (0.780)	25.2	0.69837
CYP19A1	rs143562020	Tac/Aac	Y/N	Deleterious (0.01)	Possibly Damaging (0.782)	Deleterious (0.993)	Deleterious (0.699)	24.5	0.96286
CYP1B1	rs1800440	aAc/aGc	N/S	Deleterious (0)	Possibly Damaging (0.493)	Deleterious (0.999)	Deleterious (0.609)	24.1	0.92388
CYP1B1	rs747324108	cGg/cAg	R/Q	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	34	0.95473
CYP1B1	rs9282671	Tac/Aac	Y/N	Deleterious (0)	Probably Damaging (0.997)	Deleterious (1.000)	Deleterious (0.911)	25.6	0.97751
CYP2D6	CM900081	Cca/Tca	P/S	Deleterious (0.02)	Possibly Damaging (0.89)	Deleterious (0.992)	Deleterious (0.733)	24.9	0.985
CYP2D6	rs1058172	cGc/cAc	R/H	Deleterious (0.01)	Probably Damaging (0.997)	Deleterious (1.000)	Deleterious (0.872)	35	0.99088
CYP2D6	rs143145507	aCg/aTg	T/M	Deleterious (0)	Probably Damaging (0.991)	Deleterious (1.000)	Deleterious (0.889)	25.4	0.95499
CYP2D6	rs28371703	Ctg/Atg	L/M	Deleterious (0.02)	Probably Damaging (0.977)	Deleterious (0.998)	Deleterious (0.802)	23.1	0.45911
CYP2D6	rs369177208	cGc/cAc	R/H	Deleterious (0.01)	Possibly Damaging (0.641)	Deleterious (0.991)	Deleterious (0.630)	33	0.92227
CYP2D6	rs373243894	cCc/cTc	P/L	Deleterious (0.01)	Possibly Damaging (0.815)	Deleterious (0.994)	Deleterious (0.713)	24.6	0.98527

CYP2D6	rs376636053	Tgg/Cgg	W/R	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	23.2	0.92439
CYP2D6	rs532668079	cGt/cAt	R/H	Deleterious (0.01)	Possibly Damaging (0.853)	Deleterious (0.995)	Deleterious (0.734)	32	0.98736
CYP2D6	rs59421388	Gtg/Atg	V/M	Deleterious (0)	Probably Damaging (0.987)	Deleterious (1.000)	Deleterious (0.881)	32	0.95262
CYP3A4	7:99375691	Ttt/Att	F/I	Deleterious (0.01)	Possibly Damaging (0.755)	Deleterious (0.992)	Deleterious (0.684)	24.2	0.9814
CYP3A4	rs897005504	cCt/cAt	P/H	Deleterious (0)	Probably Damaging (0.999)	Deleterious (1.000)	Deleterious (0.935)	23.7	0.90408
CYP3A4	rs12721629	Ctt/Ttt	L/F	Deleterious (0.01)	Probably Damaging (0.965)	Deleterious (0.998)	Deleterious (0.812)	24.9	0.07354
CYP3A4	rs530166880	Ctt/Ttt	L/F	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	22.6	0.17352
CYP3A4	rs540702483	Gat/Aat	D/N	Deleterious (0.01)	Benign (0.161)	Deleterious (0.989)	Neutral (0.420)	25.3	0.99234
CYP3A4	rs57409622	Cgg/Tgg	R/W	Deleterious (0)	Possibly Damaging (0.831)	Deleterious (0.999)	Deleterious (0.760)	24.4	0.06487
CYP3A4	rs67784355	aCg/aTg	T/M	Deleterious (0)	Probably Damaging (0.997)	Deleterious (1.000)	Deleterious (0.911)	23.5	0.08746
CYP3A5	7:99270249	aGa/aCa	R/T	Deleterious (0)	Possibly Damaging (0.498)	Deleterious (0.999)	Deleterious (0.611)	22.9	0.74668
CYP3A5	7:99270300	aCg/aGg	T/R	Deleterious (0)	Possibly Damaging (0.559)	Deleterious (0.999)	Deleterious (0.635)	19.58	0.72947
CYP3A5	rs142823108	aTt/aCt	I/T	Deleterious (0)	Probably Damaging (0.936)	Deleterious (1.000)	Deleterious (0.826)	27.1	0.98363
CYP3A5	rs145774441	aTt/aCt	I/T	Deleterious (0)	Possibly Damaging (0.76)	Deleterious (0.999)	Deleterious (0.726)	24.9	0.93707
CYP3A5	rs188366390	cAa/cCa	Q/P	Deleterious (0)	Probably Damaging (0.996)	Deleterious (1.000)	Deleterious (0.906)	23.6	0.91049
DPYD	COSM1688098	Gac/Tac	D/Y	Deleterious (0)	Probably Damaging (0.995)	Deleterious (1.000)	Deleterious (0.902)	26.9	0.98649
DPYD	rs114096998	Ccg/Acg	P/T	Deleterious low confidence (0)	Benign (0.415)	Deleterious (0.999)	Deleterious (0.579)	20.6	0.97773
DPYD	rs115232898	tAt/tGt	Y/C	Deleterious (0.01)	Probably Damaging (0.977)	Deleterious (0.999)	Deleterious (0.825)	25.7	0.94025
DPYD	rs142619737	Gga/Aga	G/R	Deleterious (0)	Probably Damaging (0.965)	Deleterious (1.000)	Deleterious (0.851)	32	0.95971
DPYD	rs144395748	cCt/cGt	P/R	Deleterious (0.04)	Probably Damaging (0.953)	Deleterious (0.992)	Deleterious (0.746)	28	0.99146
DPYD	rs145548112	Gca/Aca	A/T	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	34	0.98196
DPYD	rs1801158	aGt/aAt	S/N	Deleterious (0.01)	Benign (0.025)	Deleterious (0.990)	Neutral (0.407)	23.4	0.96201
DPYD	rs201268750	caC/caA	H/Q	Deleterious (0.01)	Probably Damaging (0.94)	Deleterious (0.997)	Deleterious (0.790)	27.6	0.94499
DPYD	rs2297595	Atg/Gtg	M/V	Deleterious (0)	Probably Damaging (0.985)	Deleterious (1.000)	Deleterious (0.877)	24.5	0.98591

DPYD	rs573299212	cGt/cAt	R/H	Deleterious (0)	Probably Damaging (0.977)	Deleterious (1.000)	Deleterious (0.863)	34	0.97859
DPYD	rs60511679	gTt/gGt	V/G	Deleterious (0.04)	Possibly Damaging (0.542)	Neutral (0.958)	Deleterious (0.532)	26.6	0.98805
DPYD	rs72975710	gCc/gTc	A/V	Deleterious (0)	Benign (0.168)	Deleterious (0.999)	Neutral (0.460)	25.7	0.98933
ESR1	rs188957694	cGc/cAc	R/H	Deleterious (0.01)	Possibly Damaging (0.775)	Deleterious (0.993)	Deleterious (0.694)	32	0.99659
GSTP1	rs199949339	aTt/aCt	I/T	Deleterious (0.01)	Probably Damaging (0.987)	Deleterious (0.999)	Deleterious (0.842)	27.3	0.7949
SLC19A1	rs202074305	Tac/Cac	Y/H	Deleterious (0.01)	Probably Damaging (0.989)	Deleterious (0.999)	Deleterious (0.845)	23.1	0.78266
SLC19A1	rs557998866	gAa/gGa	E/G	Deleterious (0.03)	Possibly Damaging (0.905)	Deleterious (0.989)	Deleterious (0.724)	23.5	0.05752
TYMS	rs559346114	Gcc/Acc	A/T	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	33	0.98516
XRCC1	rs138284081	Gcc/Acc	A/T	Deleterious (0.01)	Possibly Damaging (0.801)	Deleterious (0.993)	Deleterious (0.707)	34	0.85565
XRCC1	rs143917286	cCt/cTt	P/L	Deleterious (0.04)	Possibly Damaging (0.732)	Neutral (0.968)	Deleterious (0.617)	23.8	0.93527
XRCC1	rs1799782	Cgg/Tgg	R/W	Deleterious (0.02)	Possibly Damaging (0.633)	Deleterious (0.981)	Deleterious (0.604)	32	0.79285
XRCC1	rs2307177	aAt/aCt	N/T	Deleterious (0)	Benign (0.001)	Deleterious (0.999)	Neutral (0.445)	24.1	0.99228
XRCC1	rs530200061	Ggc/Agc	G/S	Deleterious (0)	Probably Damaging (1)	Deleterious (1.000)	Deleterious (0.945)	33	0.95035
XRCC1	rs532702063	Tct/Act	S/T	tolerated (0.16)	Probably Damaging (0.994)	Deleterious (0.997)	Deleterious (0.512)	24.1	0.93652
XRCC1	rs537236859	Ggc/Tgc	G/C	Deleterious (0)	Probably Damaging (0.998)	Deleterious (1.000)	Deleterious (0.919)	34	0.91856
XRCC1	rs552435216	Cgg/Tgg	R/W	Deleterious (0)	Probably Damaging (0.995)	Deleterious (1.000)	Deleterious (0.902)	34	0.66779

The 61 missense variants in Table 3.4 have the highest likelihood that the amino acid substitutions would disrupt the function of the gene. High scores in at least four of the six tools that apply to the assessment of missense variants were needed to be included in this set. The classification as high confidence variants (i.e. all tools in concordance) and the population frequencies for these variants are reported in the corresponding Table 3.5 below.

Table 3. 5 Missense variant frequencies and allele counts across African population

Gene	ID	Effect Allele	High Conf.	YRI	ESN	GWD	MSL	ZULU	LWK	BAG	ETH
ABCB1	rs137996914	A		0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs142600685	A	Yes	0 (0)	0 (0)	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs147487745	T		0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs148718120	G	Yes	0 (0)	0 (0)	0.017 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs200754866	T	Yes	0 (0)	0 (0)	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs201159898	A	Yes	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs35730308	G	Yes	0 (0)	0.005 (1)	0 (0)	0.011 (2)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs535338471	C	Yes	0 (0)	0 (0)	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs572038993	G	Yes	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
ABCB1	rs930790323	G		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.004 (1)
CYP19A1	rs142652579	T		0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP19A1	rs143562020	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.004 (1)
CYP1B1	rs1800440	C	Yes	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0.005 (1)	0.005 (1)	0.033 (8)
CYP1B1	rs747324108	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)
CYP1B1	rs9282671	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0.015 (3)	0 (0)
CYP2D6	rs1058172	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0.033 (8)
CYP2D6	rs1065852	A	Yes	0.106 (23)	0.09 (18)	0.115 (26)	0.164 (28)	0.105 (21)	0.035 (7)	0.035 (7)	0.07 (17)
CYP2D6	rs143145507	A	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (2)	0 (0)
CYP2D6	rs28371703	T		0.004 (1)	0.005 (1)	0.022 (5)	0 (0)	0.005 (1)	0.015 (3)	0 (0)	0.033 (8)
CYP2D6	rs369177208	T	Yes	0.004 (1)	0.005 (1)	0.004 (1)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)
CYP2D6	rs373243894	A	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)
CYP2D6	rs376636053	G	Yes	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
CYP2D6	rs532668079	T	Yes	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP2D6	rs59421388	T	Yes	0.106 (23)	0.08 (16)	0.137 (31)	0.094 (16)	0.12 (24)	0.171 (34)	0.22 (44)	0.02 (5)
CYP3A4	7:99375691	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)
CYP3A4	rs12721629	A		0 (0)	0.005 (1)	0 (0)	0 (0)	0.02 (4)	0.02 (4)	0.01 (2)	0 (0)
CYP3A4	rs530166880	A		0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A4	rs540702483	T		0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A4	rs57409622	A		0 (0)	0.005 (1)	0.008 (2)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A4	rs67784355	A		0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A5	7:99270249	G		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.016 (4)
CYP3A5	7:99270300	C	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)
CYP3A5	rs142823108	G	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (2)	0 (0)	0 (0)
CYP3A5	rs145774441	G	Yes	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CYP3A5	rs188366390	G	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0.04 (8)	0 (0)	0.005 (1)	0.008 (2)

CYP3A4	rs897005504	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0.012 (3)
DPYD	COSM1688098	A	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.004 (1)
DPYD	rs114096998	T		0.042 (9)	0.03 (6)	0.026 (6)	0.058 (10)	0.005 (1)	0.02 (4)	0.005 (1)	0 (0)
DPYD	rs115232898	C	Yes	0.041 (9)	0.01 (2)	0.039 (9)	0.017 (3)	0.03 (6)	0.015 (3)	0.02 (4)	0 (0)
DPYD	rs142619737	T	Yes	0.004 (1)	0 (0)	0.008 (2)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
DPYD	rs144395748	C	Yes	0 (0)	0 (0)	0.008 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
DPYD	rs145548112	T	Yes	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
DPYD	rs1801158	T		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.004 (1)
DPYD	rs201268750	T	Yes	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
DPYD	rs2297595	C	Yes	0.009 (2)	0.005 (1)	0.026 (6)	0.011 (2)	0.115 (23)	0.101 (20)	0.065 (13)	0.07 (17)
DPYD	rs573299212	T	Yes	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (2)	0 (0)	0 (0)	0 (0)
DPYD	rs60511679	C		0.009 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
DPYD	rs72975710	A		0.009 (2)	0.005 (1)	0 (0)	0 (0)	0 (0)	0.005 (1)	0.005 (1)	0 (0)
ESR1	rs188957694	A	Yes	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
GSTP1	rs199949339	C		0 (0)	0.01 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
SLC19A1	rs202074305	G		0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
SLC19A1	rs557998866	C		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0.008 (2)
TYMS	rs559346114	A	Yes	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
XRCC1	rs138284081	T		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (2)	0 (0)	0.004 (1)
XRCC1	rs143917286	A		0 (0)	0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0.01 (2)	0.012 (3)
XRCC1	rs1799782	A		0.074 (16)	0.075 (15)	0.07 (16)	0.035 (6)	0.075 (15)	0.121 (24)	0.11 (22)	0.158 (38)
XRCC1	rs2307177	G		0.041 (9)	0.05 (10)	0.057 (13)	0.023 (4)	0.015 (3)	0.03 (6)	0.095 (19)	0.041 (10)
XRCC1	rs530200061	T	Yes	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
XRCC1	rs532702063	T		0 (0)	0 (0)	0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)
XRCC1	rs537236859	A	Yes	0 (0)	0 (0)	0.004 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
XRCC1	rs552435216	A		0 (0)	0.005 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

*Number of alleles shown in parentheses

Most missense variants were detected in only one population. These variants are in most cases the result of a single heterozygote individual in the population, resulting in a very low allele frequency and thus statistical comparisons are not valuable. The missense variants for which African populations have two or more alleles (frequency ≥ 0.01) have a higher likelihood of being real (i.e. not a NGS artefact) and are more useful for PGx assessment. All these have rsIDs, which provides further confidence to their existence, even when rare.

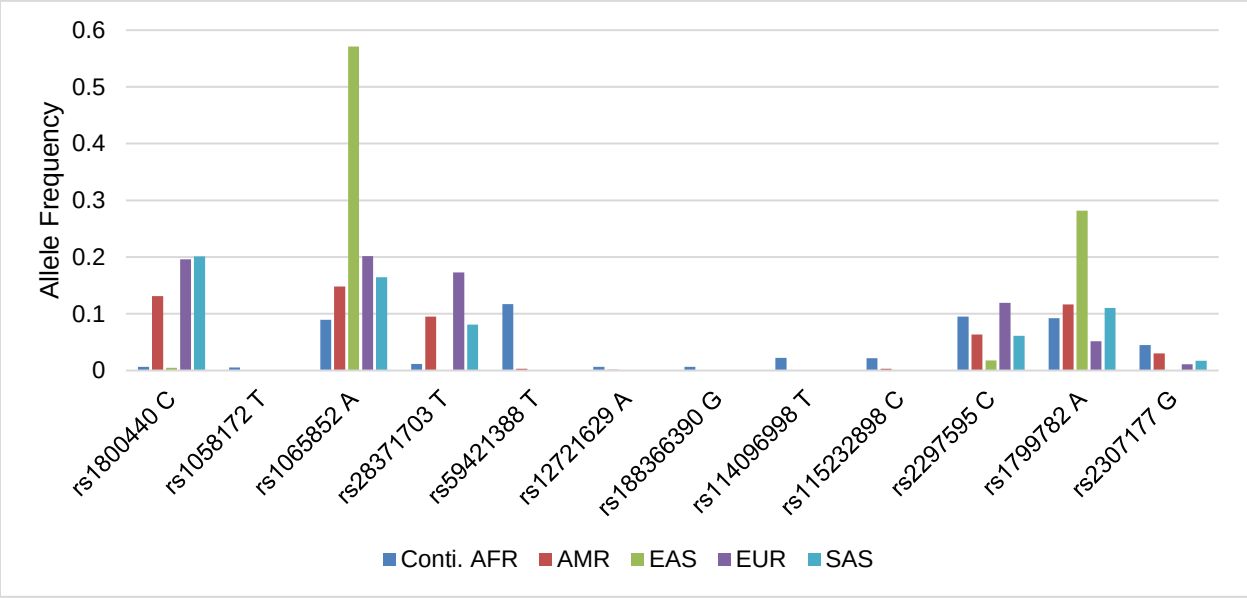


Figure 3. 12 Global frequencies of missense variants predicted to be likely functional (excluding extremely rare variants).

Most of the missense variants, though predictions indicated they are likely functional, do not have high counts of PharmGKB annotations from clinical research (with the exception of rs1065852 and rs2297595). This means that their functional relevance has not been extensively studied, or they are assumed to be less clinically relevant. This is the case even for the variants which are fairly common in several well characterised populations, and include rs1800440, rs28371703, and rs1799782. For the rare, African specific variants, such as rs114096998, and rs115232898, and the more common rs59421388 variant, the lack of clinical annotations may be due to the lack of PGx investigations in continental Africans.

3.3 PharmGKB database results for previously reported variants

A total of 31 variants were present in both PharmGKB and in the population datasets assessed in this study. Of these, only six variants, rs1801158, rs115232898, rs2297595, rs1799782, rs1065852 (missense) and rs3892097 (splice acceptor variant), were detected by the toolsets in this study (Table 3.6). The remaining 25 variants were not detected for various reasons, including that they were in intronic regions and therefore considered as low likelihood for functional impact. The summary statistics for the 33 variants including the number of PharmGKB annotations (Table 3.7), the scores the tools assigned (Table 3.8), and their respective frequencies in Africans (Table 3.9) are reported below.

Table 3. 6 PharmGKB reports for the variants detected in this study

GENE	ID	VARIANT ANNOTATION COUNT	CLINICAL ANNOTATION COUNT	LEVEL 1 OR 2 CLINICAL ANNOTATIONS
CYP2D6	rs1065852	9	2	0
CYP2D6	rs3892097*	24	5	0
DPYD	rs115232898	6	1	0
DPYD	rs1801158	13	1	0
DPYD	rs2297595	27	1	0
XRCC1	rs1799782	8	1	0

* This is the splice acceptor variant described previously in Section 3.2.2.

These six variants, though predicted likely functional and having high counts of variant annotations, do not have guidelines in place as yet that would place them into higher significance PharmGKB annotation levels. PharmGKB captures reports of drug reactions, however, circumstances around the type of association, severity of reaction and numbers of separate findings would need to be considered first before an official guideline would be appropriate.

Some of the missense variants may be relevant for PGx use in Africa, and their African allele frequencies are shown in Figure 3.13, along with global frequencies in Figure 3.14. As noted above, populations have been arranged in order of West, South, and East Africans for comparison. The splice acceptor variant in *CYP3A5* (rs3892097) is not included in the figure as frequencies are shown in Figures 3.10 and 3.11 above.

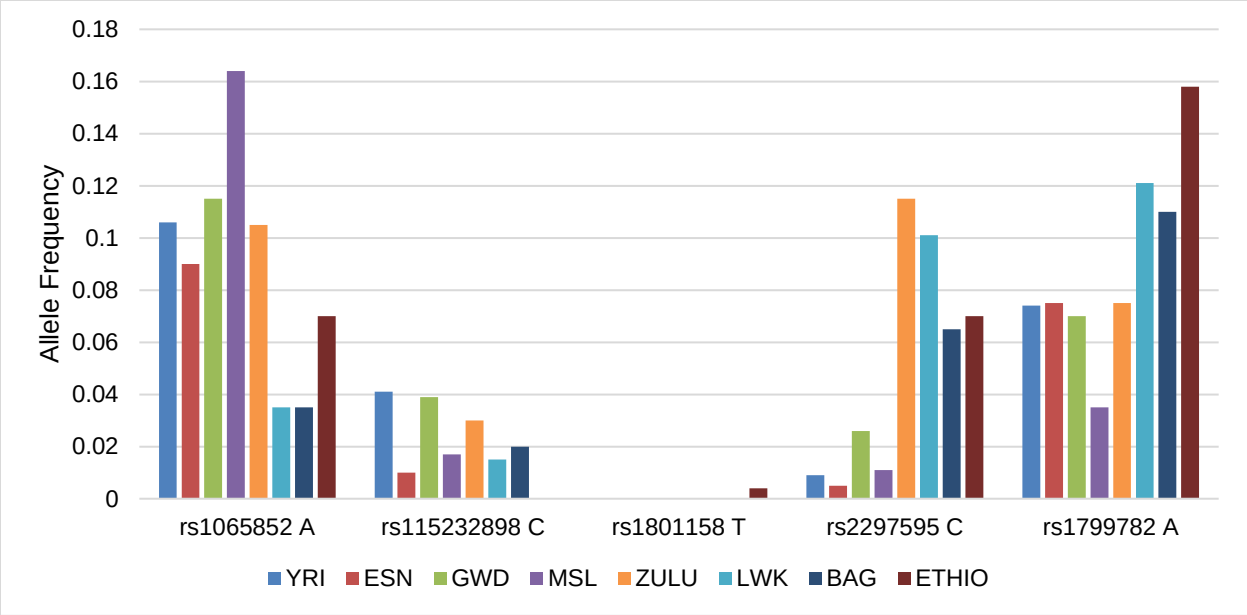


Figure 3. 13 Frequencies in African populations for variants predicted in this study to have likely functional impact and with significant annotations in PharmGKB (one variant - rs3892097, is reported above in Table 3.10)

Apart from the less common rs115232898 variant, and the rare and limited rs1801158 variant, all others show significant differences across African populations (Appendix 7.3.2). These differences are interestingly showing regional patterns. For rs1065852 the allele frequency is markedly higher in west (YRI, ESN, GWD, MSL) and South Africa (Zulu) compared to east Africa (LWK, GWD, ETHIO) For rs1799782 the frequencies are higher in east Africa and for rs2297595, east and south African populations have a higher frequency, compared to the west.

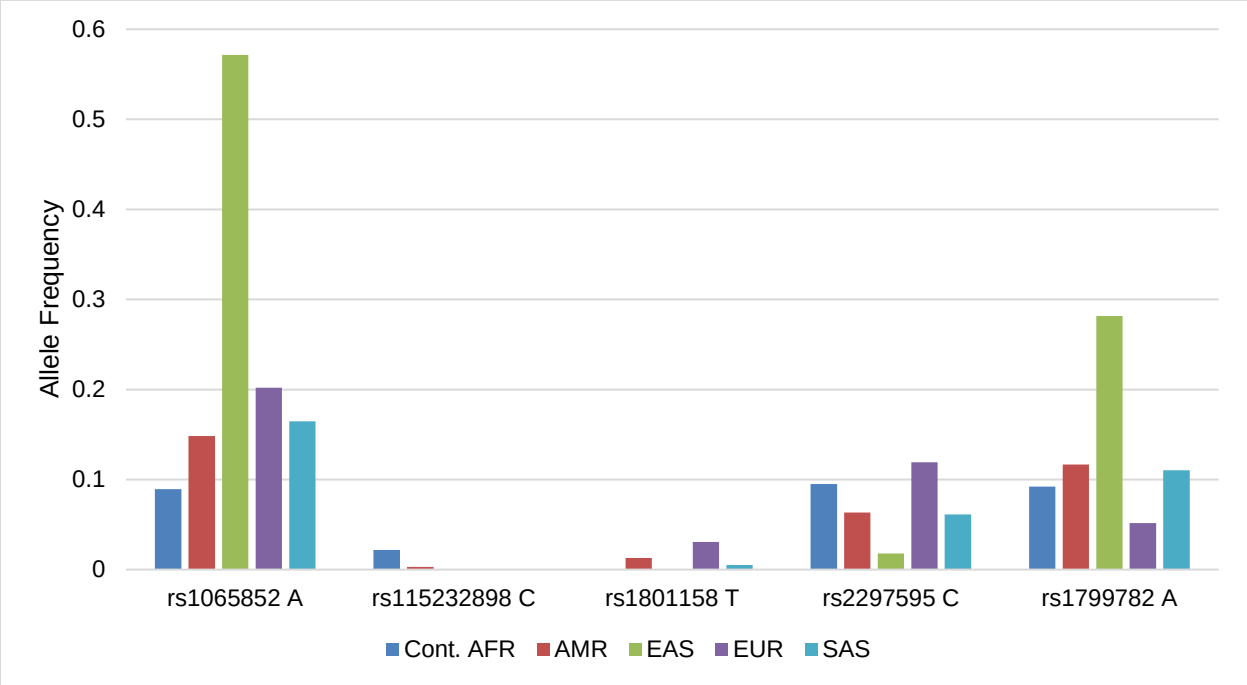


Figure 3. 14 Global frequencies for variants predicted in this study to have likely functional impact and with significant annotations in PharmGKB

The five variants (Figure 3.14) have diverse patterns of global frequency distribution, with large deviations seen even between East and South Asian populations. The allele frequencies for global-Africa are generally lower than most other populations, except for the rs115232898 variant, which is almost African specific.

In Table 3.7, the characteristics of variants that have PharmGKB annotations, but that were not detected in my study, are shown.

Table 3. 7 PharmGKB annotations counts for variants not detected via predictive scoring

GENE	ID	TYPE	VARIANT ANNOTATION COUNT	CLINICAL ANNOTATION COUNT	LEVEL 1 OR 2 CLINICAL ANNOTATIONS
ABCB1	rs1045642	Synonymous	425	80	5
ABCB1	rs1128503	Synonymous	146	25	0
ABCB1	rs2032582	Missense	227	42	2
ABCB1	rs2032583	Intron	3	1	0
ABCB1	rs2229109	Missense	17	10	0
CYP19A1	rs1008805	Intron	3	2	0
CYP19A1	rs4646	3 Prime UTR	5	4	0
CYP19A1	rs749292	Upstream	4	2	0
CYP19A1	rs6493497	Intron	3	2	0
CYP1B1	rs1056836	Missense	9	3	0
CYP2D6	rs16947	Missense	3	1	0
CYP3A4	rs2242480	Intron	17	4	0
CYP3A4	rs2740574	Upstream	57	11	1
CYP3A4	rs35599367	Intron	42	7	0
CYP3A4	rs4646440	Intron	3	1	0
CYP3A5	rs776746	Intron*****	334	32	5
DPYD	rs12022243	Intron	3	1	0
DPYD	rs17376848	Synonymous	16	1	0
DPYD	rs1801159	Missense	21	1	0
DPYD	rs1801160	Missense	16	1	0
DPYD	rs1801265	Missense	23	1	0
DPYD	rs61622928	Missense	5	0	0
ESR1	rs2234693	Intron	5	3	0
ESR1	rs9340799	Intron	8	4	0
GSTP1	rs1138272	Missense	7	2	0
GSTP1	rs1695	Missense	57	11	4
SLC19A1	rs1051266	Missense	42	3	0
TYMS	rs151264360	3 Prime UTR	21	5	1
XRCC1	rs1799782	Missense	8	1	0
XRCC1	rs25487	Missense	28	3	1
XRCC5	rs1051685	3 Prime UTR	3	1	0

The rs1045642, rs2032582, rs2740574, rs776746, rs1695 and rs25487 variants are of high relevance for PGx use (due to the Level 1 and 2 annotation records) and may be of particular interest to this study despite the fact that they were not detected as potentially significant given the assessment algorithm used in this study. Some, but not all the annotations for these variants PharmGKB are relevant to cancer treatment drugs. Despite this, it is still a strong indicator that the variant is functional, and thus may have an impact on cancer drugs depending on the nature and type of the drug. To better characterise their potential impact in Africa, comparisons of intra-African frequencies and inter-African and global frequencies are shown in Figures 3.15 and 3.16 respectively.

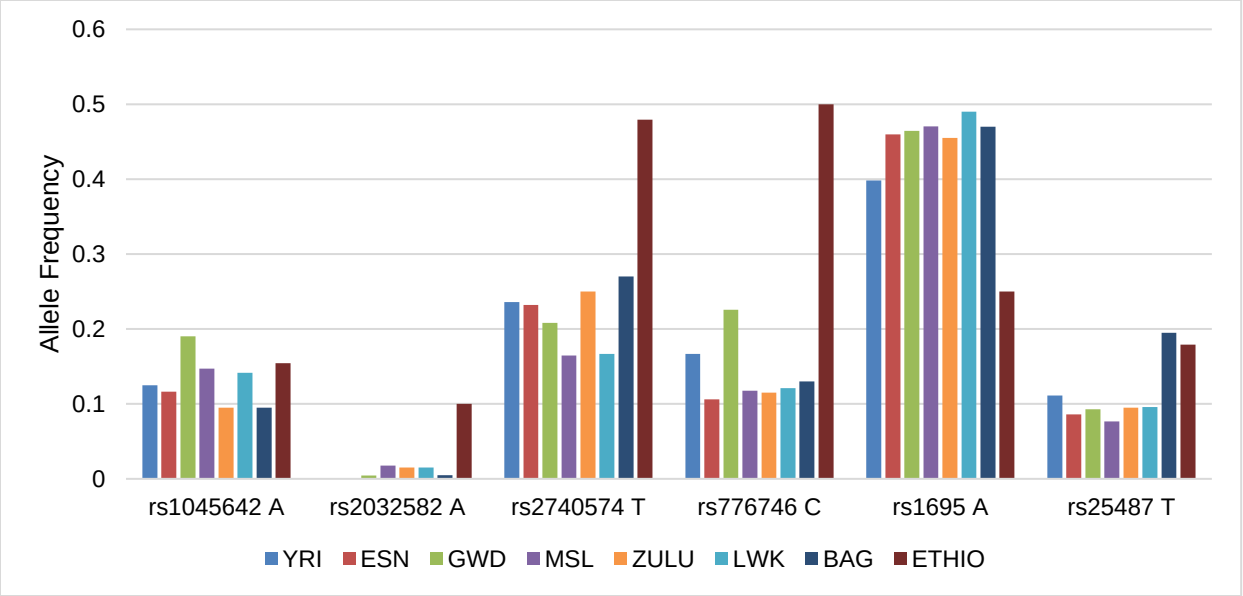


Figure 3. 15 African population frequencies for variants with high level annotations from PharmGKB

There is considerable frequency variation across African populations. For rs1045642 and rs25487 the allele frequencies range between ~0.10 and 0.20. The rs2740574 and rs776746 variants have frequencies ranging from ~0.10 - 0.26 for most populations. The rs1695 variant is more common, with frequencies ranging between ~0.40 – 0.50 for all populations except Ethiopians, which are substantially lower (0.25). For three alleles, rs2032582, rs2740574 and rs776746, the Ethiopian population has a strikingly higher allele frequency (more detail in Appendix 7.3.2). For rs104562 the Ethiopian population does not show a difference.

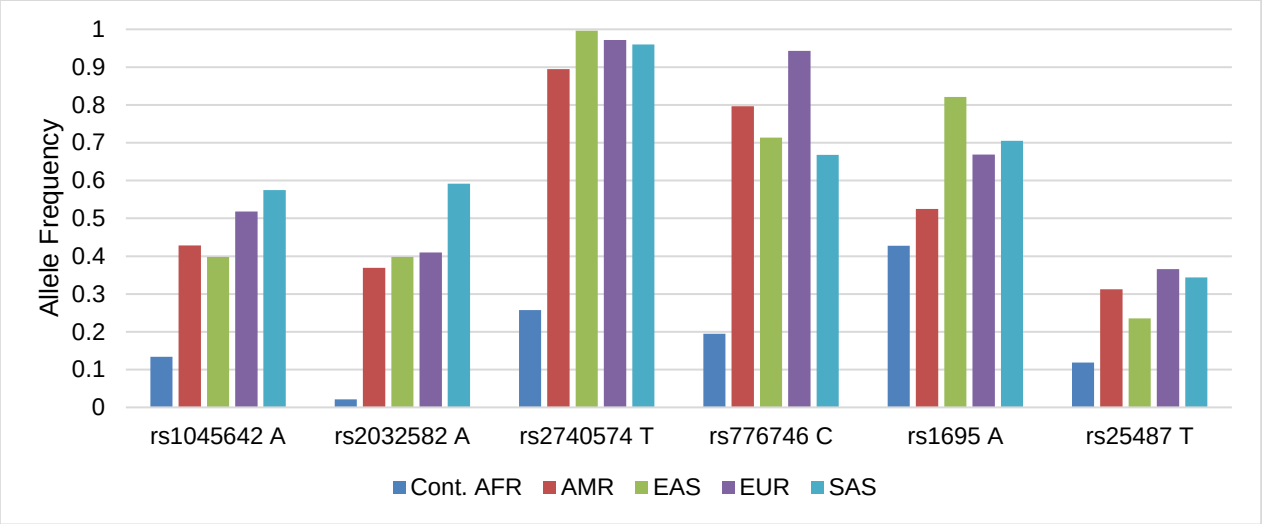


Figure 3. 16 Global allele frequencies for variants with high level annotations from PharmGKB

Frequencies for these variants which are well characterised in terms of clinical effect are significantly less common in Africans compared to the other populations. The same case was seen for the rs3892097 splice acceptor variant, as well as the for missense variants with predictions and clinical reports. This suggests that these variants would affect a smaller proportion of Africans should they be recommended for testing in a clinical setting.

To assess how the variants in Table 3.7 would have scored give the criteria described in this study, they were put through the algorithm. The scores are displayed in Table 3.8 below to characterise their predicted impact. Non-coding variants are included as CADD and FATHMM-MKL scores are used to predict the effect of these variants.

Table 3. 8 Scoring reports for variants with a significant number of Variant or Clinical annotations from PharmGKB

GENE	ID	CONSEQUENCE	SIFT	POLYPHEN-2	CAROL	CONDEL	CADD PHRED	FATHMM-MKL CODING	FATHMM-MKL NON-CODING
ABCB1	rs1045642	Synonymous	-	-	-	-	-	-	-
ABCB1	rs1128503	Synonymous	-	-	-	-	-	-	-
ABCB1	rs2032582	Missense	-	-	-	-	23.2	0.386	-
ABCB1	rs2032583	Intron	-	-	-	-	5.427	0.074	0.225
ABCB1	rs2229109	Missense	T (0.39)	B (0.002)	N (0.608)	N (0.016)	14.05	0.736	0.977
CYP19A1	rs1008805	Intron	-	-	-	-	-	-	-
CYP19A1	rs4646	3 Prime UTR	-	-	-	-	-	-	-
CYP19A1	rs6493497	Upstream	-	-	-	-	1.941	0.004	0.086
CYP19A1	rs749292	Intron	-	-	-	-	3.149	0.000	0.046
CYP1B1	rs1056836	Missense	T (0.34)	B (0.026)	N (0.642)	N (0.024)	14.46	0.180	0.337
CYP2D6	rs16947	Missense	T (0.44)	B (0)	N (0.560)	N (0.013)	0.042	0.002	0.129
CYP3A4	rs2242480	Intron	-	-	-	-	-	-	-
CYP3A4	rs2740574	Upstream	-	-	-	-	5.335	0.029	0.168
CYP3A4	rs35599367	Intron	-	-	-	-	2.187	0.012	0.090
CYP3A4	rs4646440	Intron	-	-	-	-	1.885	0.050	0.159
CYP3A5	rs776746	Intron*****	-	-	-	-	-	-	-
DPYD	rs12022243	Intron	-	-	-	-	1.245	$5x10^{-4}$	0.059
DPYD	rs17376848	Synonymous	-	-	-	-	0.005	0.522	0.937
DPYD	rs1801159	Missense	D (0.04)	B (0)	N (0.960)	N (0.349)	9.639	0.808	0.954
DPYD	rs1801160	Missense	T (0.1)	B (0.298)	N (0.882)	N (0.378)	25.9	0.980	0.996
DPYD	rs1801265	Missense	T (0.52)	B (0)	N (0.009)	N (0.009)	-	-	-
DPYD	rs61622928	Missense	T (0.31)	B (0.002)	N (0.688)	N (0.027)	17.28	0.885	0.987
ESR1	rs2234693	Intron	-	-	-	-	-	-	--
ESR1	rs9340799	Intron	-	-	-	-	-	-	-
GSTP1	rs1138272	Missense	T (0.14)	B (0.065)	N (0.848)	N (0.254)	13.65	0.190	0.215
GSTP1	rs1695	Missense	T (1)	B (0)	N (0.000)	N (0.000)	0.001	0.034	-
SLC19A1	rs1051266	Missense	T (0.35)	B (0)	N (0.650)	N (0.021)	0.025	0.065	0.788
TYMS	rs151264360	3 Prime UTR	-	-	-	-	-	-	-
XRCC1	rs25487	Missense	T (0.6)	B (0)	N (0.004)	N (0.004)	9.117	0.262	-
XRCC5	rs1051685	3 Prime UTR	-	-	-	-	-	-	-

*****The rs776746 variant, listed as intronic, is only so in the canonical transcript, but causes a splice acceptor variant in common, albeit non-canonical transcripts.

Table key: T – Tolerated; D – Deleterious; B – Benign; N – Neutral. These are estimates of the variant impact based on the tool score as defined by the authors.

For non-coding variants above that have CADD and FATHMM MKL scores, none reach the functional significance thresholds assigned by the developers of these toolsets. Some variants are missing scores; this is a result of the toolset not having a score available for that genomic reference site in the pre-computed datasets used for this annotation pipeline. The *ABCB1* rs2032582 missense variant was not scored using SIFT and Polyphen-2, and thus no result is produced for CAROL and Condel. The rs2032582 variant is tri-allelic, with the A and C alleles common worldwide, with the minor allele of the three being T, which is a known pathogenic variant (T is absent in continental Africans). These scores were not retrieved in the VEP pipeline used, which may be due to it having A or C as reference matches in the reference file used.

The African frequencies for variants that have PharmGKB annotations, but that were not detected in my study, are shown in Table 3.9 below.

Table 3. 9 Frequencies in African populations of variants with annotations from PharmGKB

Gene	ID	Allele	YRI	ESN	GWD	MSL	ZULU	LWK	BAG	ETH
ABCB1	rs1045642	A	0.125 (27)	0.116 (25)	0.19 (41)	0.147 (32)	0.095 (21)	0.141 (31)	0.095 (21)	0.154 (33)
ABCB1	rs1128503	A	0.144 (31)	0.101 (22)	0.164 (35)	0.135 (29)	0.1 (22)	0.111 (24)	0.095 (21)	0.163 (35)
ABCB1	rs2032582	A	0 (0)	0 (0)	0.004 (1)	0.018 (4)	0.015 (3)	0.015 (3)	0.005 (1)	0.2 (43)
ABCB1	rs2032583	G	0.241 (52)	0.217 (47)	0.168 (36)	0.135 (29)	0.205 (44)	0.258 (56)	0.24 (52)	0.225 (49)
ABCB1	rs2229109	T	0 (0)	0 (0)	0.004 (1)	0 (0)	0 (0)	0.006 (1)	0 (0)	0.021 (4)
CYP19A1	rs1008805	G	0.125 (27)	0.136 (29)	0.084 (18)	0.135 (29)	0.105 (23)	0.177 (38)	0.18 (39)	0.313 (68)
CYP19A1	rs4646	A	0.333 (72)	0.283 (61)	0.204 (44)	0.282 (61)	0.36 (78)	0.293 (63)	0.325 (70)	0.308 (67)
CYP19A1	rs6493497	A	0.259 (56)	0.303 (65)	0.261 (56)	0.241 (52)	0.385 (83)	0.303 (65)	0.33 (71)	0.308 (67)
CYP19A1	rs749292	A	0.495 (107)	0.485 (105)	0.46 (99)	0.406 (88)	0.475 (103)	0.419 (91)	0.36 (78)	0.342 (74)
CYP1B1	rs1056836	G	0.125 (27)	0.162 (35)	0.177 (38)	0.135 (29)	0.205 (44)	0.212 (46)	0.215 (46)	0.396 (85)
CYP2D6	rs16947	G	0.44 (95)	0.434 (94)	0.465 (100)	0.388 (84)	0.47 (102)	0.354 (76)	0.27 (58)	0.408 (88)
CYP3A4	rs2242480	C	0.153 (33)	0.091 (20)	0.177 (38)	0.106 (23)	0.125 (27)	0.091 (20)	0.17 (37)	0.483 (104)
CYP3A4	rs2740574	T	0.236 (51)	0.232 (50)	0.208 (45)	0.165 (36)	0.25 (54)	0.167 (36)	0.27 (58)	0.479 (104)
CYP3A4	rs35599367	A	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.013 (3)
CYP3A4	rs4646440	A	0.088 (19)	0.111 (24)	0.04 (9)	0.029 (6)	0.09 (19)	0.101 (22)	0.145 (31)	0.021 (4)
CYP3A5	rs776746	C	0.167 (36)	0.106 (23)	0.226 (49)	0.118 (25)	0.115 (25)	0.121 (26)	0.13 (28)	0.5 (108)
DPYD	rs12022243	T	0.204 (44)	0.141 (31)	0.15 (32)	0.118 (25)	0.115 (25)	0.202 (44)	0.16 (35)	0.196 (42)
DPYD	rs17376848	G	0 (0)	0.02 (4)	0.009 (2)	0.012 (3)	0 (0)	0 (0)	0.01 (2)	0.004 (1)
DPYD	rs1801159	C	0.148 (32)	0.187 (40)	0.08 (17)	0.071 (15)	0.185 (40)	0.303 (65)	0.225 (49)	0.163 (35)
DPYD	rs1801160	T	0.014 (3)	0.005 (1)	0.04 (9)	0.012 (3)	0.01 (2)	0.061 (13)	0.04 (9)	0.067 (14)

DPYD	rs1801265	G	0.435 (94)	0.429 (93)	0.465 (100)	0.394 (85)	0.385 (83)	0.5 (108)	0.48 (104)	0.338 (73)
DPYD	rs61622928	T	0.097 (21)	0.066 (14)	0.111 (24)	0.076 (17)	0.075 (16)	0.056 (12)	0.125 (27)	0.108 (23)
ESR1	rs2234693	T	0.472 (102)	0.414 (89)	0.447 (97)	0.412 (89)	0.27 (58)	0.383 (83)	0.385 (83)	0.379 (82)
ESR1	rs9340799	G	0.241 (52)	0.303 (65)	0.177 (38)	0.271 (58)	0.365 (79)	0.288 (62)	0.36 (78)	0.275 (59)
GSTP1	rs1138272	T	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.015 (3)	0.005 (1)	0.013 (3)
GSTP1	rs1695	A	0.398 (86)	0.46 (99)	0.465 (100)	0.471 (102)	0.455 (98)	0.49 (106)	0.47 (102)	0.25 (54)
SLC19A1	rs1051266	C	0.329 (71)	0.384 (83)	0.261 (56)	0.306 (66)	0.235 (51)	0.308 (67)	0.35 (76)	0.292 (63)
TYMS	rs151264360	TTAA AG	0.426 (92)	0.465 (100)	0.425 (92)	0.406 (88)	0.355 (77)	0.379 (82)	0.365 (79)	0.333 (72)
XRCC1	rs25487	T	0.111 (24)	0.086 (19)	0.093 (20)	0.076 (17)	0.095 (21)	0.096 (21)	0.195 (42)	0.179 (39)
XRCC5	rs1051685	G	0.329 (71)	0.313 (68)	0.372 (80)	0.394 (85)	0.36 (78)	0.308 (67)	0.35 (76)	0.196 (42)

*Number of alleles shown in parentheses

3.4 Common Variants

This section highlights variants that are common to most African populations, organized according to gene. Since they would affect more individuals in Africa, they should be assessed further to determine whether they may be useful to guide PGx strategies and treatment for public health systems.

Common variants that were either predicted to have an impact on gene function, or those that are known to have PGx implications according to PharmGKB records are displayed in the figures below. These compare African populations to one another, and then show how the average continental African frequency compares to that of other global population groups. The p-values above the variants in the comparison of African frequencies are used to show whether African frequencies differ from each other in a grouped comparison (see Methodology 2.3). Pairwise comparisons of African populations are available in Appendix 7.3.2, which can be used as a reference to compare African populations individually. Continuing the arrangement as in the Figures in preceding sections, African populations have been arranged in order of West, South, and East Africans for comparison.

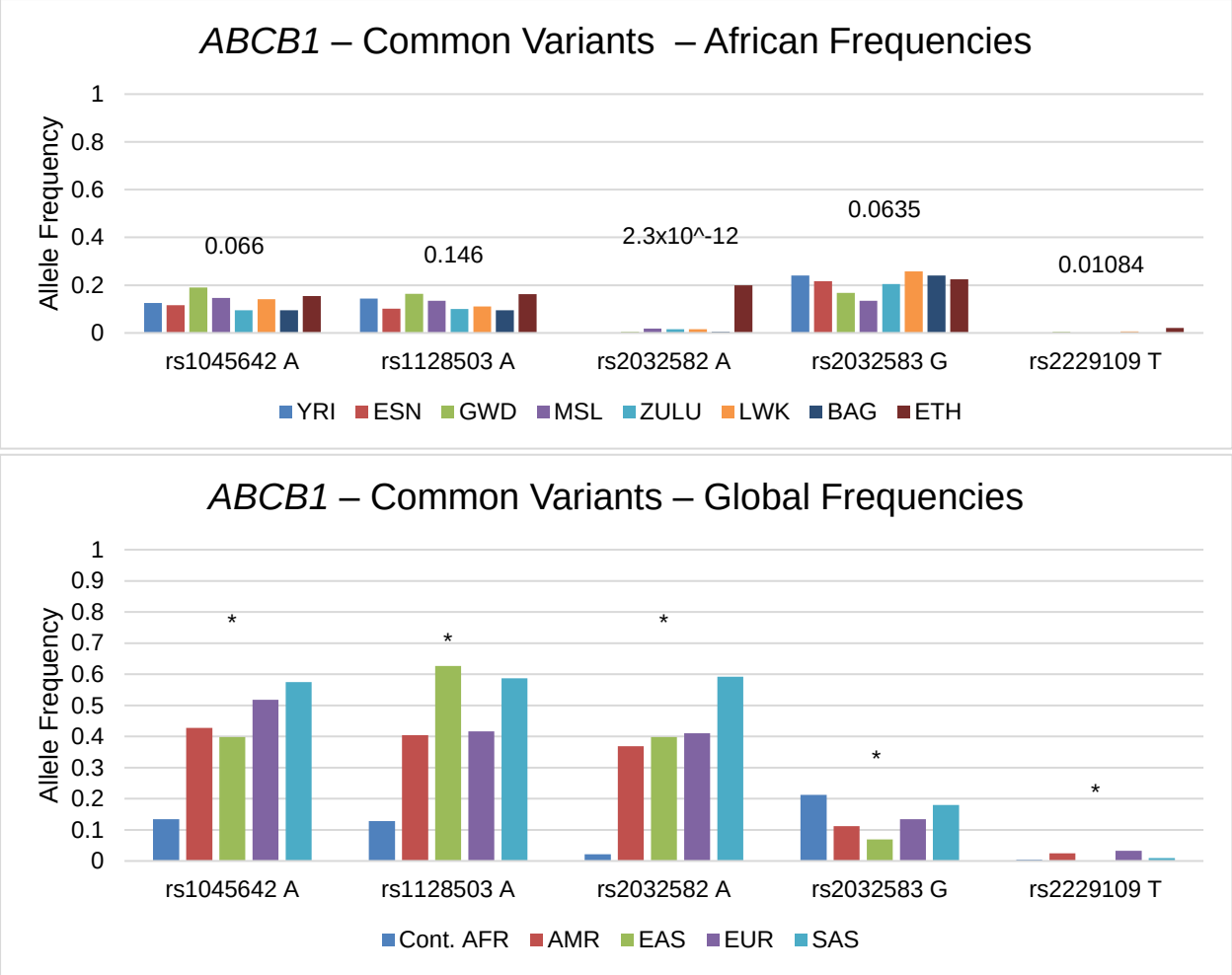


Figure 3. 17 Frequency of common variants in the *ABCB1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

The *ABCB1* gene encodes a key drug transporter and had five common variants of interest (Fig. 3.17). Most *ABCB1* variants are significantly less common in Africans than in other global populations, with the exception of rs2032583. The rs1045642 and rs1128503 variants all have high numbers of Clinical Annotations and are all found in relatively equal distribution across Africans. Only rs2032583 is commonly distributed across African populations with higher frequencies in continental Africans than the other global groups. Interestingly, the rs2032582 variant is common in Ethiopians but extremely rare in all other African populations. South Asians have high frequencies of these variants and are only surpassed by East Asians for rs1128503. The rs2229109 T allele is uncommon globally, and only appears in the Gambian, Luhya and Ethiopian populations, at extremely low frequencies.

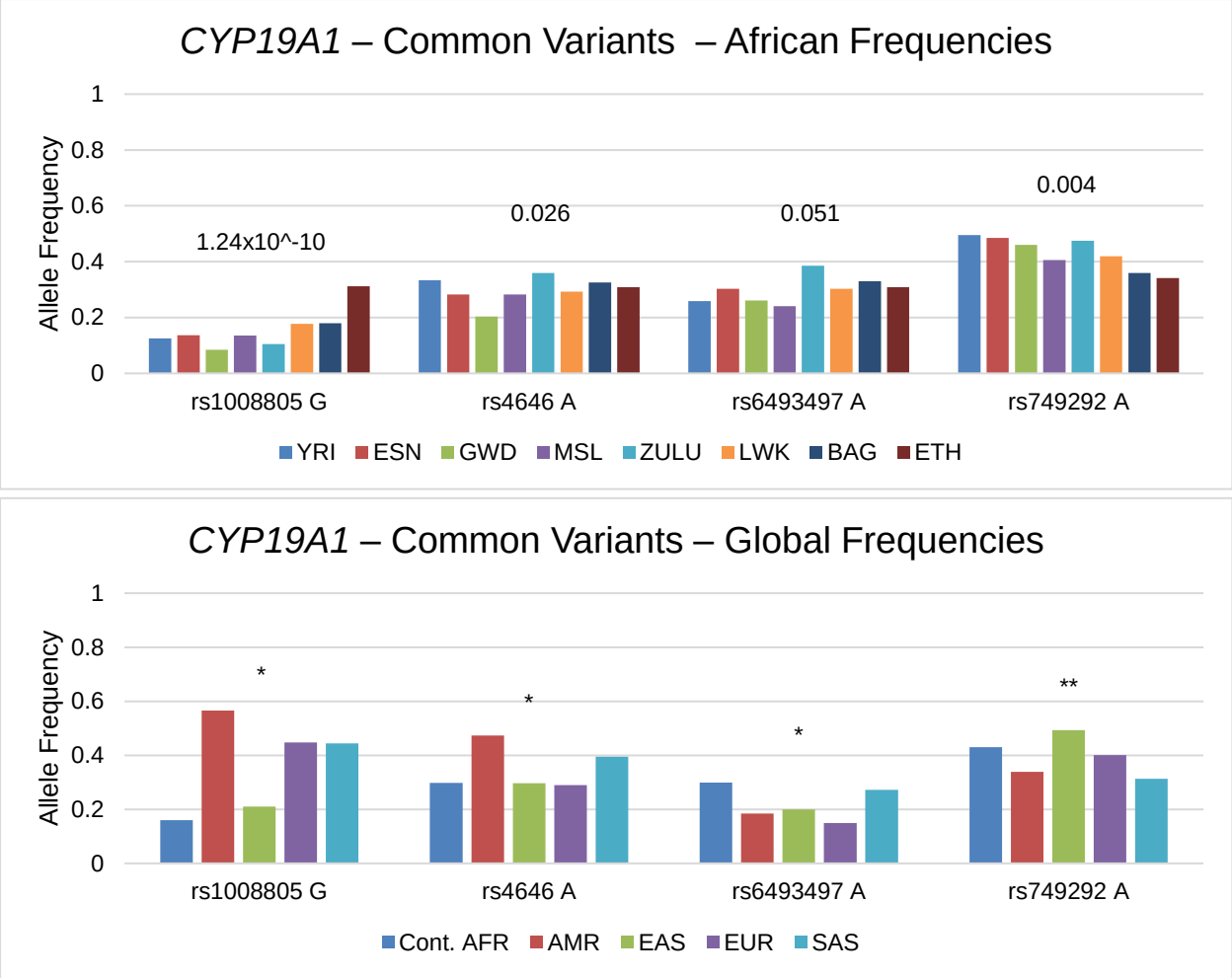


Figure 3. 18 Frequency of common variants in the *CYP19A1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

Four variants were common in the human aromatase gene (Fig. 3.18). The rs1008805 is significantly different between Africans due it being more common in Ethiopians, which have significantly different frequencies to all other African populations (Appendix 7.3.2). This variant is more common in other global populations than Africans. The rs4646, rs6493497, and rs749292 variants are not significantly different between Africans, and are present at high allele frequencies, which are similar to those in global populations. The rs749292 variant is only not significantly different between Africans and Europeans.

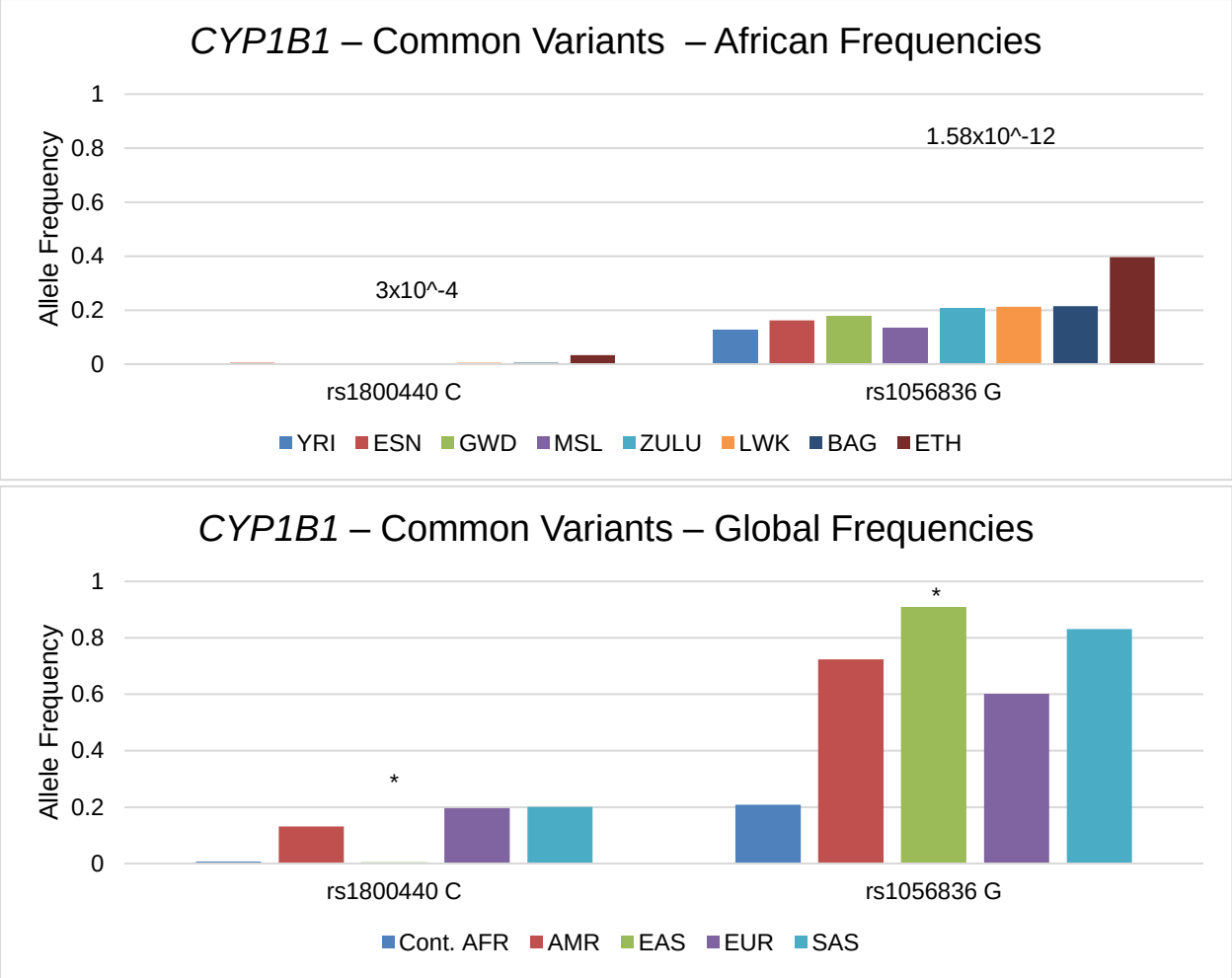


Figure 3. 19 Frequency of common variants in the *CYP1B1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

Only two common variants were present in the *CYP1B1* metabolic enzyme gene (Fig. 3.19). The rs1800440 variant is only present in Ethiopians and is interestingly also nearly absent in East Asians. Ethiopians also account for the significant difference in frequency between Africans for rs1056836, as the frequency in Ethiopians is double to that in the other African populations. The rs1056836 is far more common globally and is approaching fixation in East Asians.

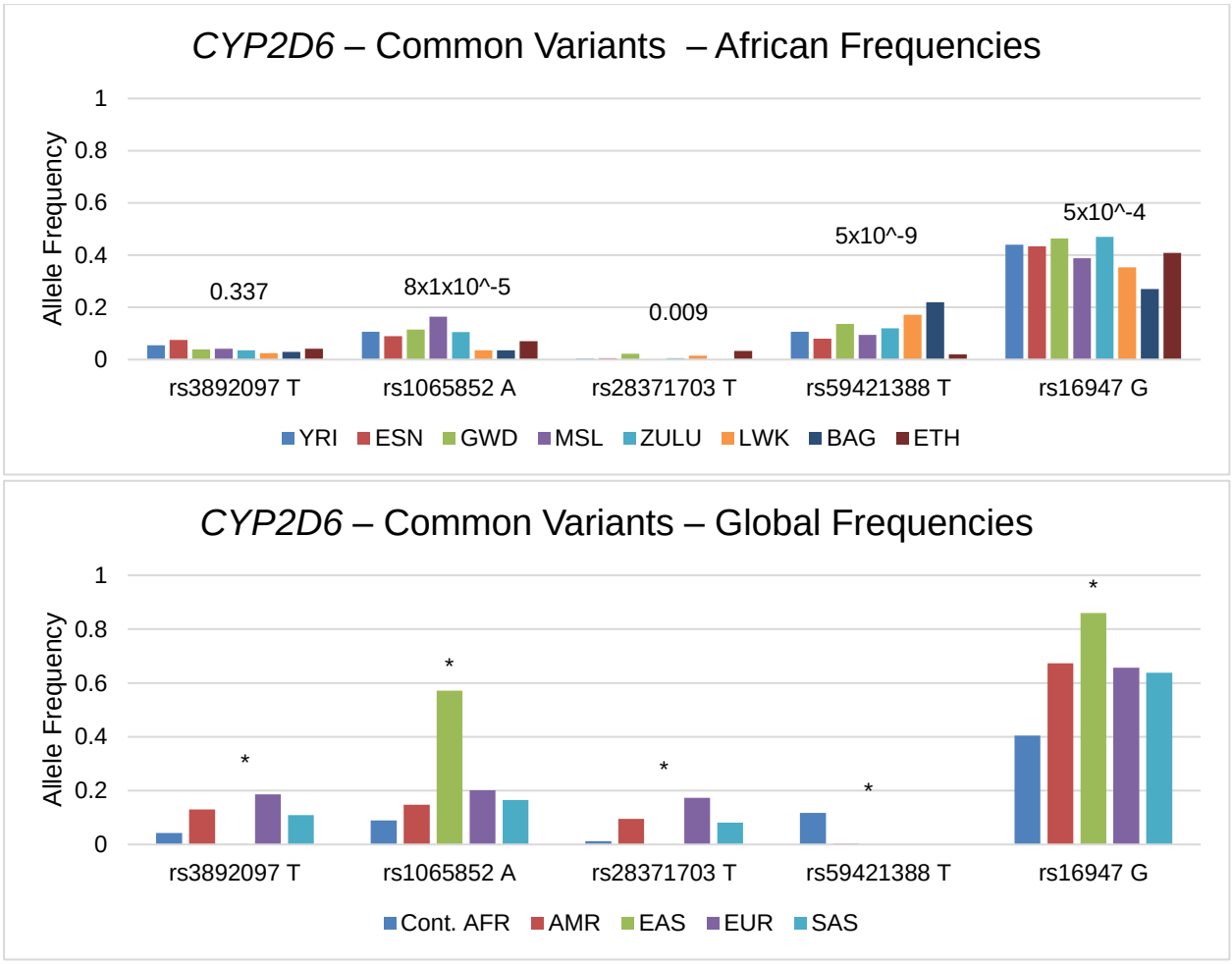


Figure 3. 20 Frequency of common variants in the *CYP2D6* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

There were five common variants in the key drug metaboliser *CYP2D6* (Fig. 3.20). The rs3892097 splice acceptor variant is present across African populations at a low frequency which is consistent across populations. This variant is more common in other global populations and is at a frequency of ~20% in Europeans. It is a variant with known PGx impact as it has multiple Variant and Clinical Annotations within PharmGKB. The other common variants detected in this gene are missense variants, with only rs1065852 and rs16947 having PGx related reports in PharmGKB. The rs1065852 variant is not evenly disturbed across African populations and appears to be present more in West and South Africans than East Africans. All West African populations have significantly distinct frequencies to that in East African populations (Appendix 7.3.2). The rs28371703 variant is evenly uncommon across Africans, but more common in Europeans. The rs59421388 variant is notably African specific,

being found only in African populations although very rarely in Ethiopians. This is interesting as the other East African populations, Bagandan and Luyha, display the highest frequencies for the variant. The rs16947 has a few annotations recorded in PharmGKB but was predicted to be likely non-functional via bioinformatics methods. This variant is common in Africa, although less so in Bagandans. The global frequencies of this variant are much greater than those of continental Africans.

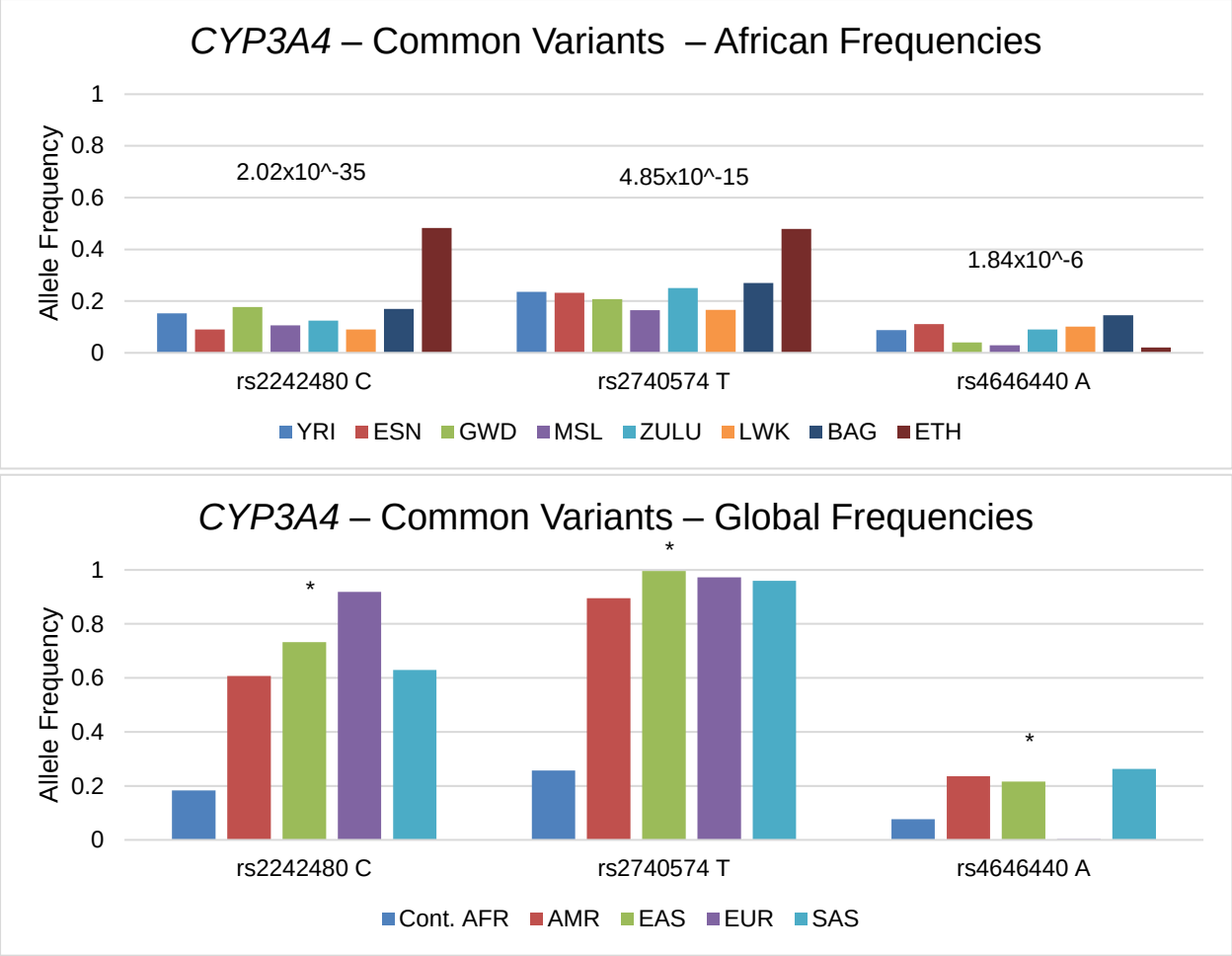


Figure 3. 21 Frequency of common variants in the *CYP3A4* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

Three common variants were in the *CYP3A4* drug metaboliser gene (Fig. 3.21). These variants were all assessed due to PharmGKB reports – two intronic variants, rs2242480 and rs4646440, and rs2740574, an upstream gene variant. These are all less common in African populations than the other global groups, particularly rs27405474, which are near fixation in non-African populations for this variant. The rs2242480 and rs27405474 variants are most common in Ethiopians but is more evenly distributed in other African populations.

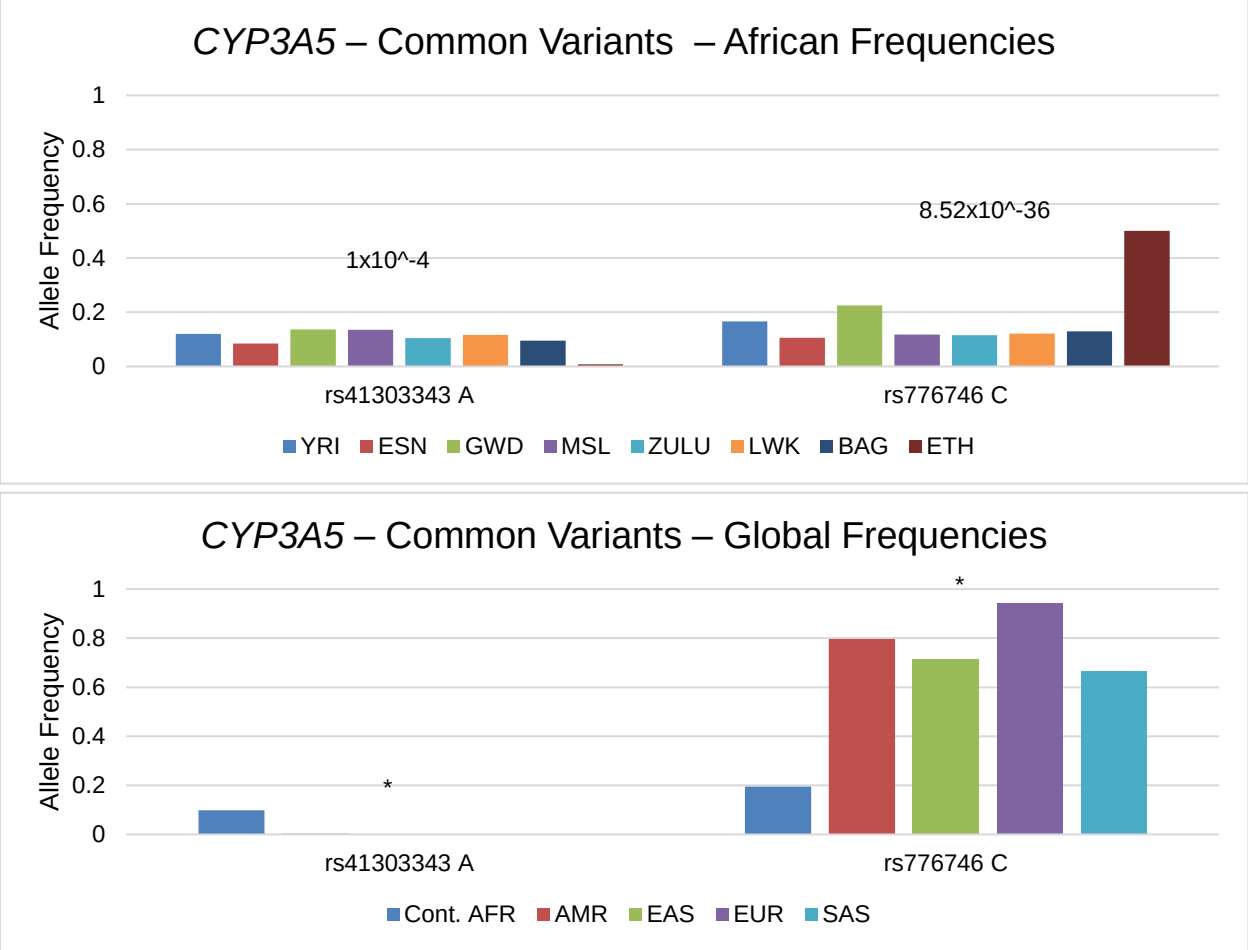


Figure 3. 22 Frequency of common variants in the *CYP3A5* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

Two interesting common variants were detected in the *CYP3A5* drug metaboliser gene (Fig. 3.22). The rs41303343 variant is particularly interesting as it is an African specific LoF type variant causing a frameshift error. It is distributed across most African populations but is rare in Ethiopians. The rs776746 variant, though an intronic variant in the canonical transcript, is a well described splice acceptor variant which is less common in Africans, but very common in other global populations with Europeans approaching fixation. This variant is more common in Ethiopians, which account for the greatest difference in frequency for African populations. These patterns indicate that the gene is more likely to have a disrupted, or non-functional state in non-African populations.

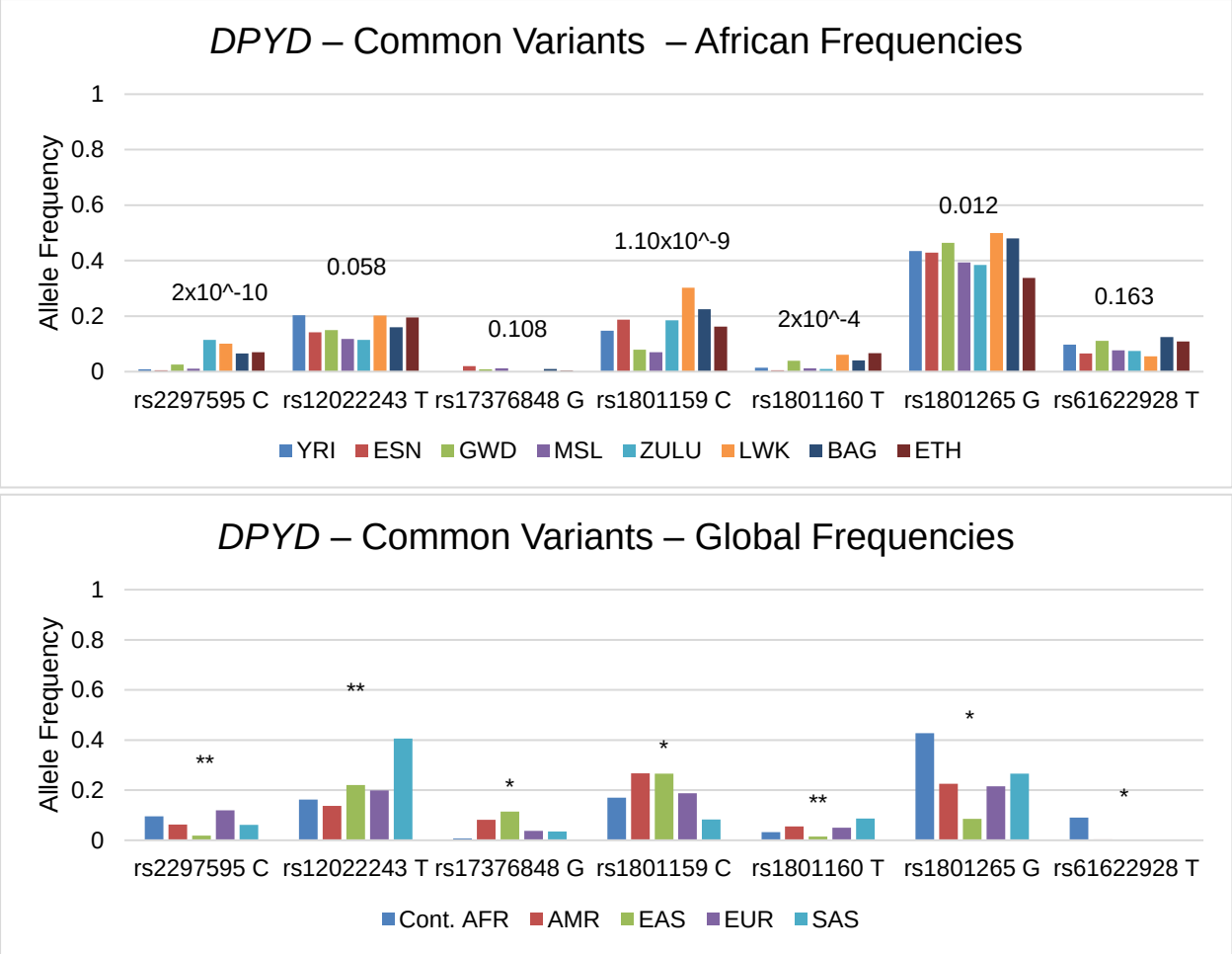


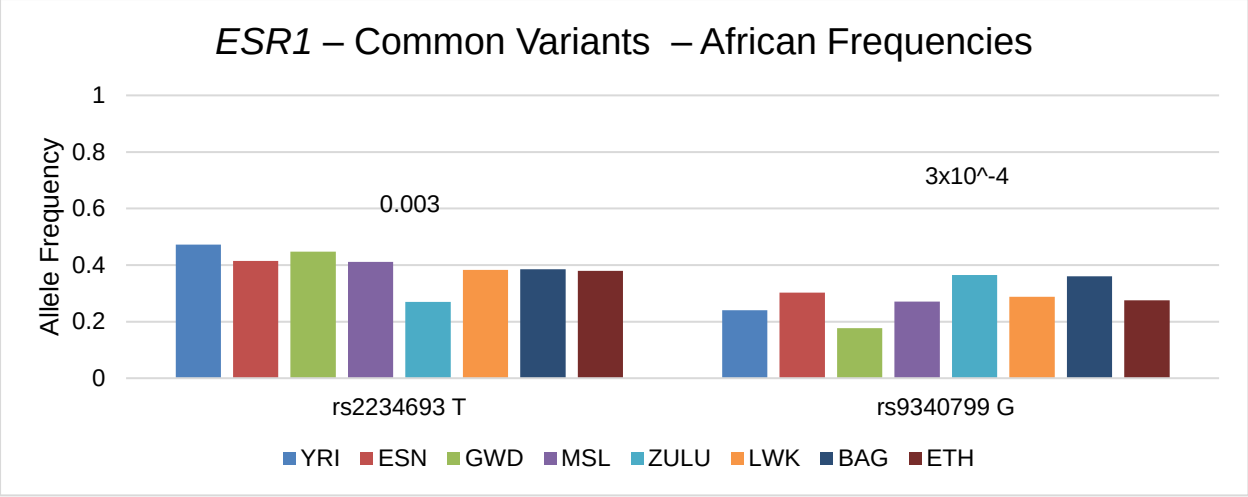
Figure 3. 23 Frequency of common variants in the *DPYD* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

DPYD has the greatest number of common variants detected of all genes assessed in this study (Fig. 3.23). The majority of these are missense variants, except for rs12022243 and rs17376848, which are intronic and synonymous respectively. The rs12022243 variant is evenly distributed across Africans, and ranges in similar frequencies in other global populations, except for South Asians where the variant is twice as common. The rs17376848 variant is uncommon in continental Africans compared to other populations.

The rs2297585 missense variant was predicted to be likely functional and has substantial Variant Annotations recorded with the PharmGKB database. This variant is more common in East and South Africans than West Africans. It is also present globally, and only East Asian frequencies are significantly different ($p=3.60x10^{-14}$) from those in continental Africans. The

rs115232898 variant, (Figures 3.13-3.14 above), is uncommon in Africans but absent in other global populations. This variant was omitted from Figure 3.23 for clarity but is present in Africans at allele frequencies between 1-4% and has recorded PharmGKB annotations. The other four *DPYD* missense variants were not detected via prediction methods (due to low scores) but have recorded findings in PharmGKB. The rs1801159 variant differs in frequency across Africans but not in distinct geographic patterns. The variant is more common in East and South Africans, than overall West Africans, but the Yoruba and Esan Nigerian populations have substantially higher frequencies than their Gambian and Mende counterparts. This variant is also common globally, though less so in South Asians. The rs1801160 variant is uncommon across Africans, but with relatively higher frequencies in East Africans. It is also relatively more common in the other global population groups, except in East Asians, where frequencies are significantly lower than in continental Africans. The rs1801265 variant is very interesting, it is very common in Africans, with an even distribution across populations. Globally, this variant is more common in Africans than all other populations. The rs61622928 variant, though not as common as rs1801265, is however specific to continental Africans. It is evenly present across African populations, and with some Variant Annotations reported, is a key variant to investigate further in Africans as it is not present in other well characterised populations.



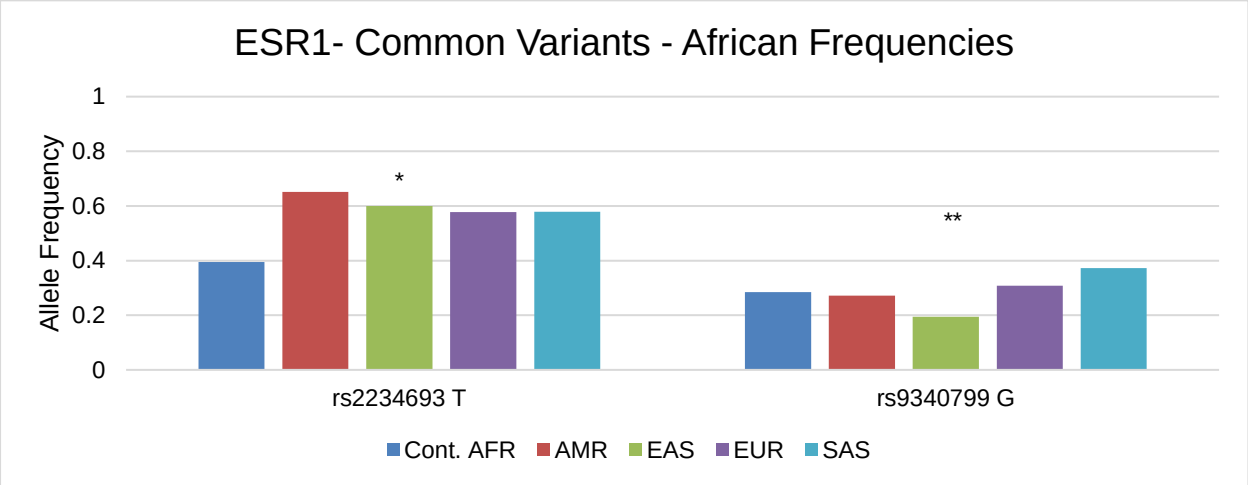


Figure 3. 24 Frequency of common variants in the *ESR1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

Two common intronic variants were found to have numerous Variant and Clinical Annotations recorded within PharmGKB within the *ESR1* gene (Fig. 3.24). The rs2234693 variant is common to all African populations and is slightly more common in West African populations, but less so in South African Zulu. This variant is significantly more common in all other global populations than continental Africans. The rs9340799 variant is present in all African populations though not in even frequency distributions. The frequencies are also not aligned in geographical patterns, with Esan, Mende, Luhya and Ethiopian populations sharing similar allele frequencies in the range of 27-30%, while the variant is less common in Gambians (17%). Global frequencies are only distinct from continental Africans in East and South Asian populations.

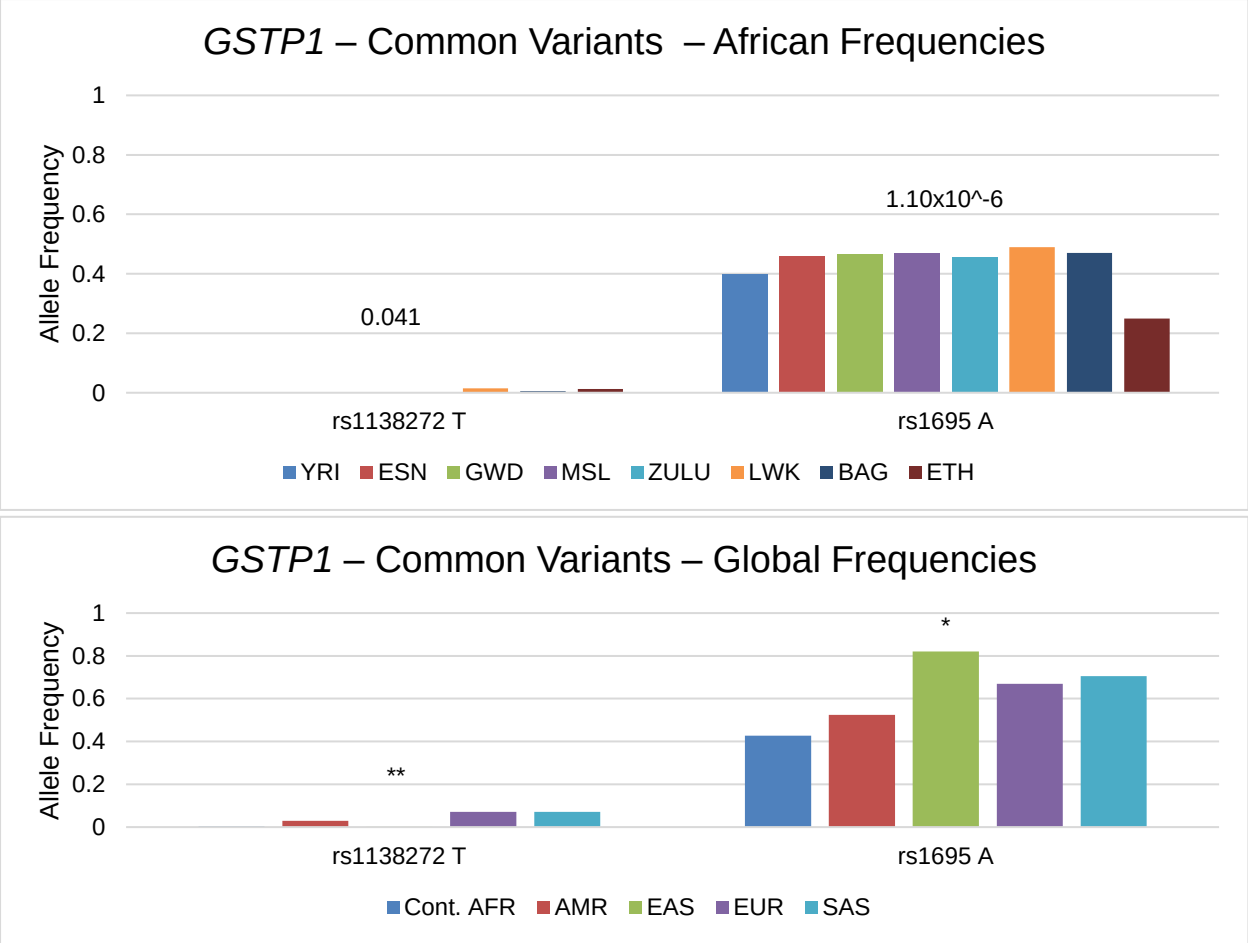


Figure 3. 25 Frequency of common variants in the *GSTP1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

Two common *GSTP1* missense variants had PharmGKB annotation records despite low scoring predictions (Fig. 3.25). The rs1138272 variant is very rare in Africans, found only East Africans at extremely low counts (1-3). It is more common in Europeans and South Asians at an allele frequency of 7%. The rs1695 variant is more relevant as it is very common (~40%) in Africans, present evenly in all populations but Ethiopians, where it is only half as common. This variant is more common in all other global populations than in continental Africans.

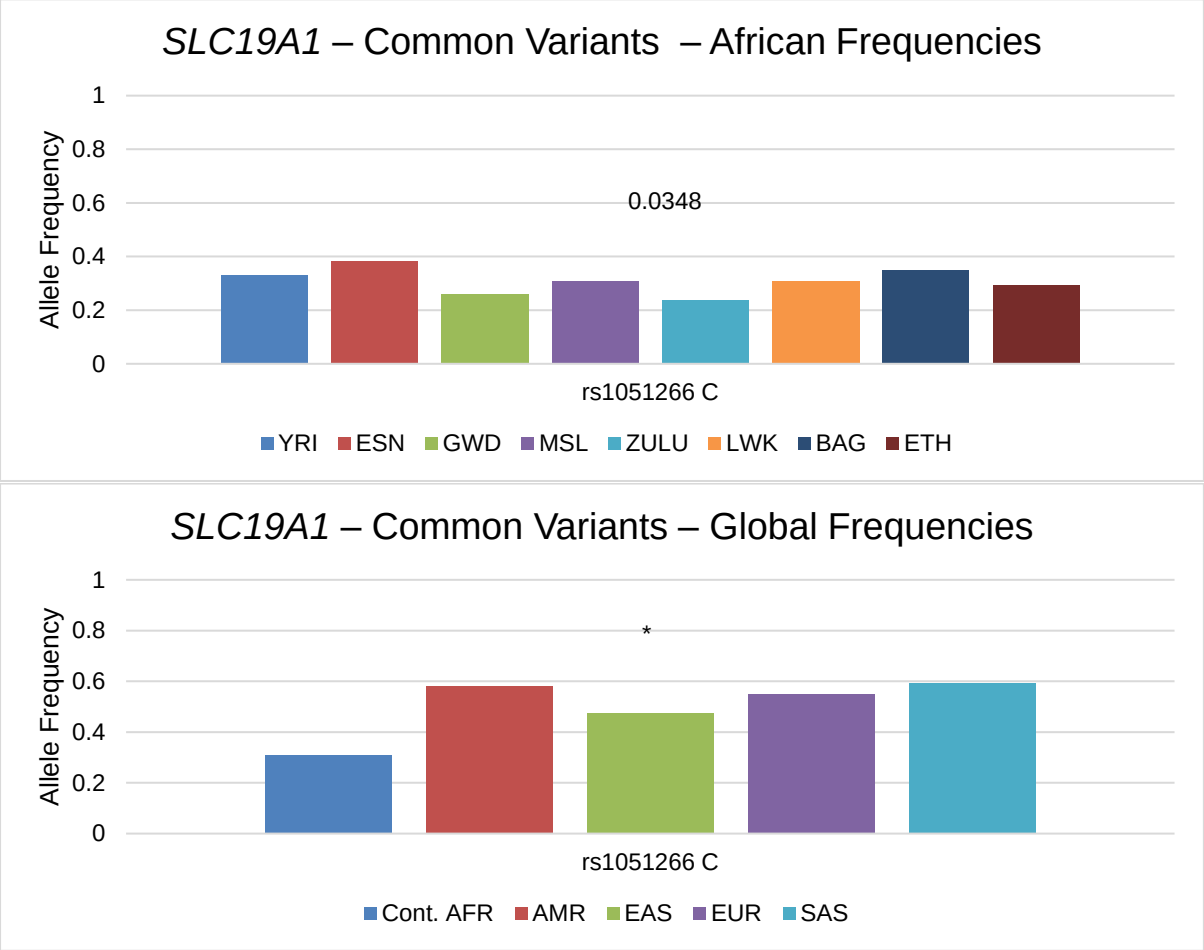
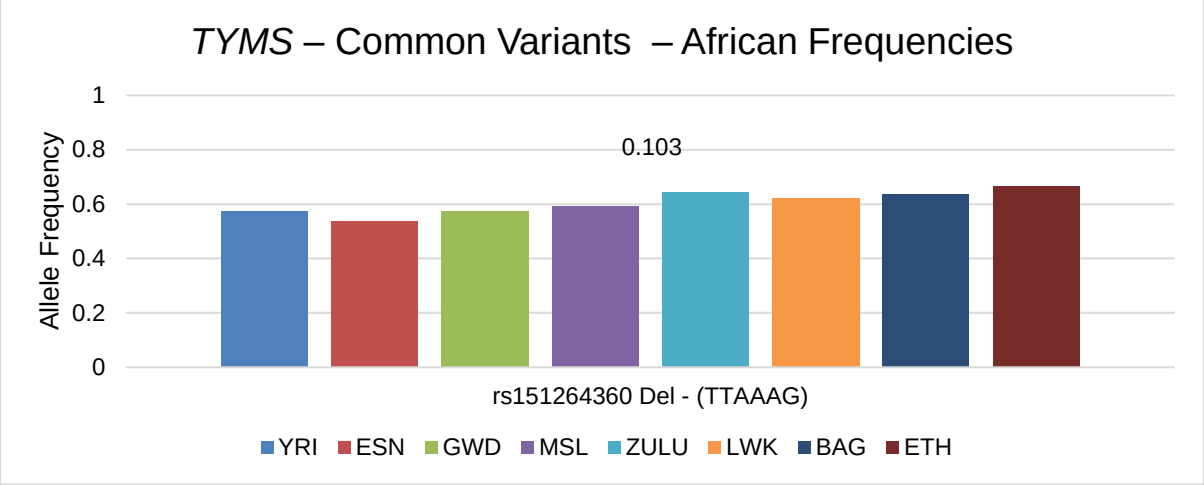


Figure 3. 26 Frequency of common variants in the *SLC19A1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

The rs1051266 is the only common missense variant within the *SLC19A1* gene assessed, and despite low prediction scores has many Variant Annotations within PharmGKB (Fig. 3.26). It is common and evenly distributed across all African populations. As with most variants with reported PGx impact, this variant less common in continental Africans than non-African populations.



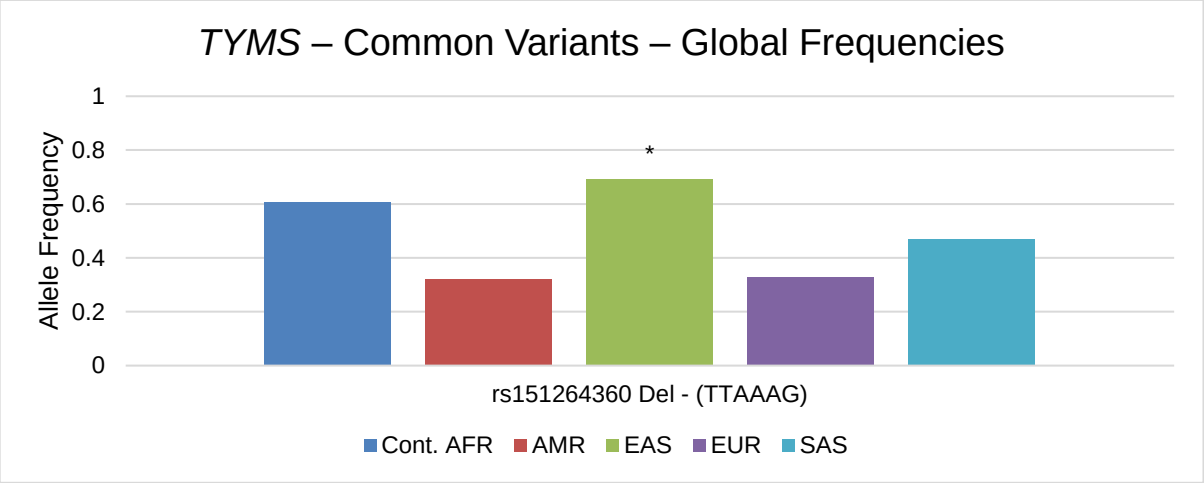


Figure 3. 27 Frequency of common variants in the *TYMS* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

The rs151264360 variant is a 3' UTR region variant characterising a six base-pair deletion (Fig. 3.27). The frequencies displayed above are for the deletion, not the alternative allele, which is TTAAAG. The deletion is common in all African populations and is evenly distributed across them. It is more common in continental African and East Asian populations than the other global population groups

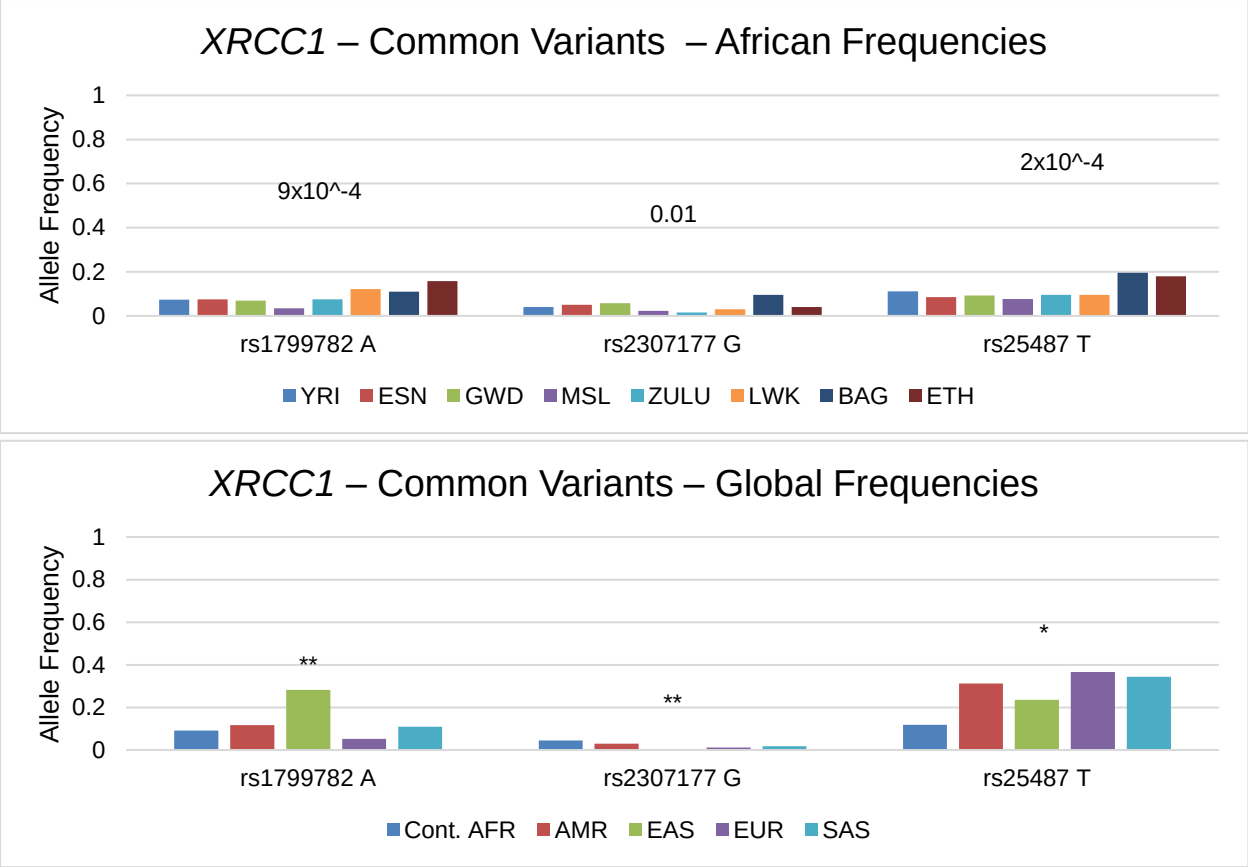


Figure 3. 28 Frequency of common variants in the *XRCC1* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

The *XRCC1* gene (one of two DNA repair protein genes assessed) had three common variants of interest assessed (Fig. 3.28). Two variants, rs1799782 and rs2307177 were predicted to be likely functional, with the former also having Variant Annotations recorded in PharmGKB. The rs1788782 variant is present in all African populations but is more common in East African populations. This variant is significantly more common in East and South Asians, and less common in Europeans than in continental Africans. Admixed American frequencies are not significantly different from continental Africans. The rs2307177 variant is uncommon in Africans and although frequencies are evenly distributed across all eight populations together, significant intra-African variability is present between the upper frequencies ~10 % in Bagandans, and the lower frequencies ~2% in Mende and Zulu. Although with an average frequency of ~5%, this variant is most common in continental Africans compared to other global populations. The rs25487 variant is a missense variant with multiple PharmGKB reports but low functional prediction scores. It is present in all African populations, though it is more

common in Bagandan and Ethiopian populations. It is significantly less common in continental Africans than in other global populations.

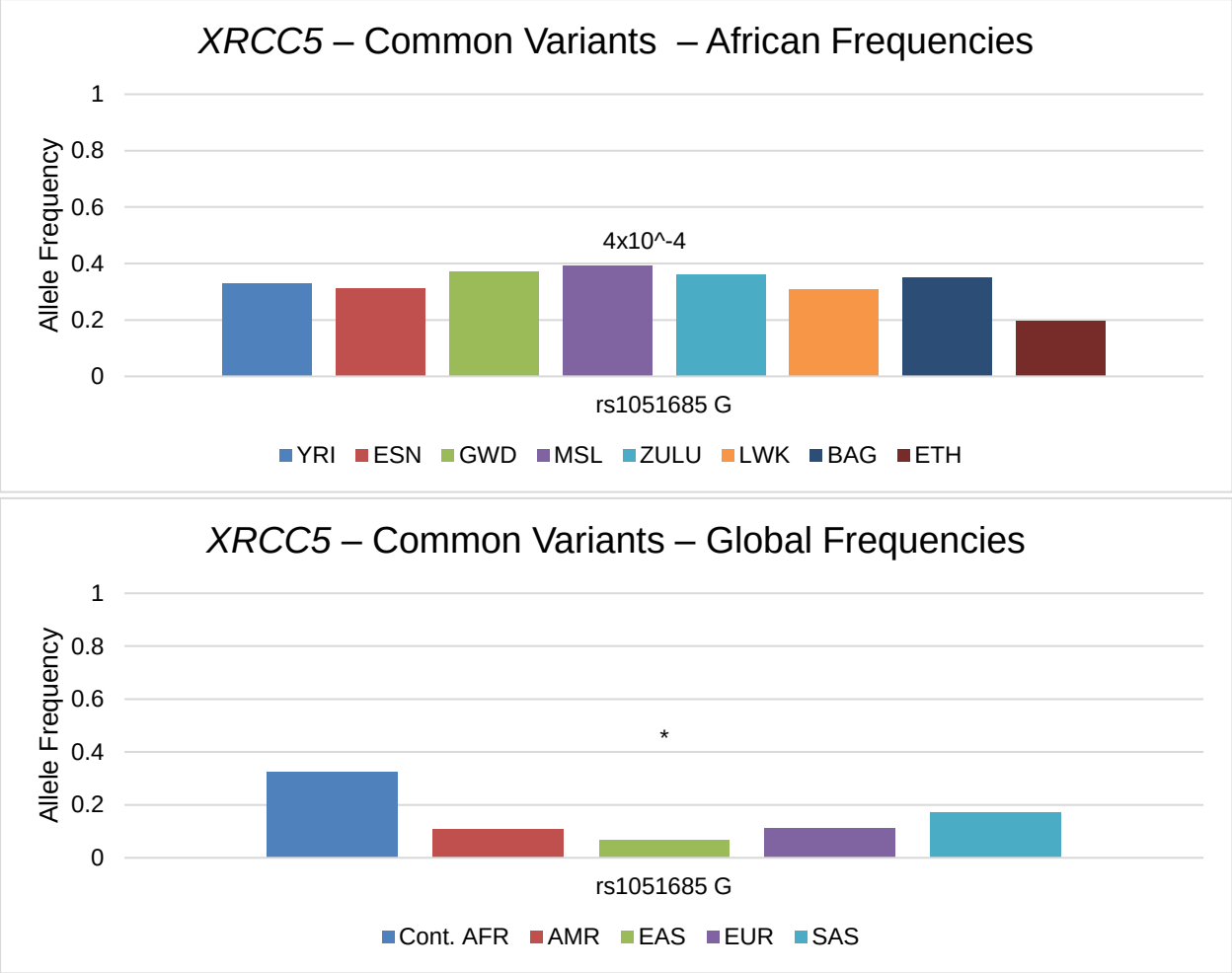


Figure 3. 29 Frequency of common variants in the *XRCC5* gene for African and Global populations

For African Frequencies, the numbers indicated above the bars are p-values for differences in alleles for the grouped African populations (described in Methodology 2.3). For Global Frequencies, a single asterisk "*" above the above the bars indicate all global populations are significantly different, while a double asterisk "**" indicates that one or more (but not all) are significantly different.

The *XRCC5* gene had only one variant of interest assessed, due to PharmGKB records (Fig. 3.29). The rs1051685 variant is a 3' UTR variant, which is common in all African populations assessed but less so in Ethiopians. Frequencies for this variant do not follow geographic patterns in Africans, as no significant deviation appears between populations from East and West Africa. It is significantly more common in continental Africans than in other global population groups.

Chapter 4: Discussion

Cancer rates are rising rapidly in Africa, and over time, these patients will need effective and safe treatment. Some genetic variants are known to induce ADRs and/or poor drug response, and prior knowledge of these variants can be used to guide treatment to ensure safety and efficacy for the patient. These strategies would be highly effective to address the rising rates of cancer in Africa, however, PGx knowledge of cancer treatment is currently underrepresented in African populations. The aim of this study was to annotate genetic variation in selected genes related to metabolism, transport and targets of cancer drugs in continental African populations, and to compare the frequency of these variants in Africans to other global populations. The annotation was performed on selected gene datasets for eight African populations, extracted from publicly available whole genome sequence datasets from the KGP and the AGVP. Selection of genes was based on a series of previously reported genes that impact cancer treatment for the most common cancer types predicted to rapidly increase in Africa. The selection criteria included genes that are relevant to several cancer types and are also relevant to the largest variety of drugs known to have PGx based interactions. The annotation of variants in each of the genes was performed with the bioinformatics tool VEP, along with a series of additional plugins that can predict if a variant has a likely deleterious functional impact on the gene product. Annotations and scores for variants (and the specific allele of these variants) were generated and if considered important in terms of PGx allele frequency comparisons were made between populations.

A database of PGx related variants, obtained from PharmGKB, was mined for variants within the selected genes with previously reported clinical effects on drug reactions. The PharmGKB variant list and the annotated variant list from this study were compared and examined for variants which were present in both. In total, 101 variants of interest were found to be likely involved in PGx reactions from both functional prediction and PharmGKB records. For these variants, frequency comparisons were done for populations within Africa and globally to examine significant differences that could impact the choice of cancer related drugs used for a specific population. All variants likely to impact function, from both the bioinformatics evaluation and the previously reported findings, were compared among the African populations as well as against global populations.

Where possible variant impact was interpreted in the context of clinical findings, but it was often found that data were contradictory or severely limited. There was also very little data

from African populations relating variants to performance in specific clinical trials. Variants identified as likely functional were discussed in terms of their predicted impact (which is correlated to available clinical findings,) and their relative frequency, both within Africans, and between Africans and other global populations. This created an understanding to promote more robust conclusions when assessing variants for likely pharmacogenomic impact and to determine whether there are significant differences in Africans compared to other populations. This information could motivate and guide future research initiatives in Africa, specifically for cancer treatments.

4.1 Characterisation of variants and their potential functional impact

All the African populations studied had roughly equal numbers of variants across all 13 genes tested. The median range of ~13 000 variants for each population showed no apparent regional differences. There were some gene specific differences, in terms of more variability in the number of variants per gene between the populations. For example, in *DPYD* the Gambian (GWD) population has 480 variants more than Mende (MSL). This observed population difference is mainly due to the number of intronic variants, while coding sequence and regulatory variants are more similar in number across populations. Very few LoF variants were observed across the WGS of 824 African individuals, and most were extremely rare and was often observed in a single population. Most LoF type variants were present in the *CYP3A5* and *CYP2D6* genes. With the exception of rs3892097 (*CYP2D6*) and rs41303343 (*CYP3A5*), all other LoF variants are of extremely low frequency, and most limited to a single population group. These variants are often not conserved across populations due to their knock-out effect on genes (MacArthur *et al*., 2012).

As can be expected, the total numbers of variants within genes are directly related to their size. The *ESR1* and *DPYD* genes are the largest genes analysed, and so contain the majority of all variants assessed. The number of intronic variants is also directly related to the size of the gene, as the largest genes (along with larger intron size) have the greatest number of intronic variants present. There is significant overlap of intronic variation between the populations, which was assessed as the number of unique intronic variants across all populations relative to the total count of intronic variants per population. With the *ABCB1* gene as an example, there are an average of ~1000 variants in each population in this gene, meaning that if they were all different, there would be a total of 8000 variants. However, *ABCB1* has only ~3000 unique variants in a non-redundant count of all variants assessed across populations, indicating a substantial degree of variant overlap across populations.

Despite the size of *ESR1,* there are fewer coding variants compared to other large genes. Only one *ESR1* variant was found to be likely deleterious. This is in contrast to other larger genes, such as *ABCB1* and *DPYD*, which contain the highest counts of likely deleterious missense variants assessed in this analysis. Though missense variants are less common in *ESR1,* distal regulatory factors are known to drive its oncogenic effect (Bailey et al., 2016). This may suggest that though sequence conservation is higher in the gene coding regions, variants in regulatory regions are less restricted. The variants assessed by Bailey *et al.* 2016 lie up to 5000 bases upstream from transcription, and were not captured in this assessment which only used a 1000bp window.

Missense variants were relatively abundant and when analysed per gene, showed some interesting differences in the proportion of variants that were predicted as likely deleterious to gene function. Proportions of those predicted to be deleterious range from ~10% (*GSTP1* and *ESR1*) to ~55% (*CYP3A4* and *CYP3A5*). These genes had similar numbers of overall missense variants (9 to13 missense variants), thus a gene with more missense variants does not necessarily have more variants predicted to have deleterious function. This may be due to factors related to enzyme flexibility. *GSTP1* variants in flexible regions have been shown to increase function (Stenberg *et al.*, 2000), while CYP P450 flexible regions are key to enabling substrate promiscuity (Otyepka *et al.*, 2012), and this may be more conserved in regards to function. The genes with the greatest number of missense variants, i.e. *ABCB1* (28 variants), *DPYD* (28 variants) and *CYP2D6* (25 variants), have similar proportions of likely deleterious variants at 36%, 39%, and 36% respectively.

It is interesting to note the differences in the number of upstream and downstream gene region variants between genes, despite the fact that each assesses a fixed 1000bp region. *CYP1B1*, *ESR1*, and *XRCC1* have a greater number of upstream variants, indicating that these genes may be influenced more by regulatory related effects than others. Of downstream variants, it is interesting that *SLC19A1* has a much greater number of variants than any other gene assessed, indicating less conservation in the region following the end of transcription for this gene.

4.1.1 Non-coding and regulatory region variants with high likelihood scores

The potential functions of the non-coding and regulatory variants are difficult to interpret in terms of PGx due to their indirect relationship to protein function. Most PGx research to date is conducted with exome sequencing and therefore coding regions, potentially missing key non-coding variant effects. A caveat in this study was that only two tools were suitable to assess

these variant types, and in order to reduce false positives, additional comparisons to PharmGKB reported variant were used. None of the variants scoring high using these tools have previous reports in PharmGKB. While this may be due to a lack of published data on these variant types, the tools failed to detect variants with significant numbers of previous annotations. This places doubt on the ability of these tools to accurately predict PGx related effects for non-coding and regulatory variants.

It is also noted that genes that have larger numbers of overall variants yielded more variants in this category. With *DPYD* and *ESR1* as an example, these genes have the greatest numbers of high scoring intronic variants, while also having the greatest number of intronic variants overall. Similarly, *SLC19A1*, which yielded high scoring downstream variants, had the highest number of downstream gene variants overall. Though the potential of these tools to assess variants is well described (Richardson *et al*., 2016), without more information to confirm the potential effect of these variants, it emphasises that the use of only two tools is insufficient to assess these variants.

4.1.2 Non-coding and regulatory region variants from PharmGKB with reported clinical effects

As variants which were predicted to be deleterious computationally were cross-examined against the PharmGKB database, it was clear that some variants which have clinical findings attached were not detected using the bioinformatics tools used in this assessment. In PharmGKB there are several non-coding variants with multiple annotations attached. These were not detected using the scoring method developed to assess the likely functional impact in this study. Low scores in these cases could be explained by numerous factors involving the variant type and tool design, and as such, are discussed in regard to each case variant in section 4.2. Intronic and intergenic variants were excluded from assessment using coding variant prediction tools. For variants with clinical significance which have potential regulatory effects, all were excluded by not meeting score cut-offs for CADD and FATHMM, the two toolsets which can score non-coding variants. Some of these variants have missing scores for these tools, because they are absent in the reference datasets (Kircher *et al.*, 2014; Shihab *et al.*, 2015) from which the scores were generated.

4.1.3 Coding SNVs with clinical effects, that were not detected using the annotation algorithms in this study

For the missense variants that have previously reported clinical effects but were not detected using the annotation algorithm developed in this study, almost all had very low score

indications for being deleterious, particularly when using SIFT and Polyphen-2. This was most likely due to the amino acid substitution being for a highly similar amino acid. Amino acids which are similar in size, charge and polarity are often not likely to disrupt the structure of the entire protein, thus SIFT will not score these changes as likely deleterious (Ng and Henikoff, 2003). Polyphen-2 will also score these particular variants in this manner, but will also assess if they are present in highly conserved sites (Shihab *et al.*, 2013). Some examples are the *DPYD* rs1801265 variant and the *CYP2D6* rs16947 variant, which although having low predictive scores, have been strongly associated with clinical effects for specific drugs. In these cases, it may be that the variant is co-segregating with another functional variant or that the variant causes disruptions in drug interaction mechanisms.

An exception is the *ABCB1* rs2032582 variant, where a missense variant for which VEP did not return SIFT and Polyphen-2 scores (likely due to it being tri-allelic) has numerous clinical annotations. The variant allele A, which is present in Africans at very low frequencies, is common in other world populations, along with the C allele. The third and minor allele T, is not present in continental Africans but is a known pathogenic variant, which induces increased risk of developing lung cancer (Gervasini *et al.*, 2006).

In addition, synonymous variants would not be scored as deleterious as several of the tools used evaluate amino acid differences and their impact on the protein. It is therefore not surprising that variants in the excluded regions and synonymous variants that had previously been associated with drug effects were missed.

4.2 Impact of genetic variation on Drug Transporters

Among the 13 genes analysed two encoded important drug transport proteins. They are *ABCB1* and *SLC19A1*, and play key roles in the transport of a variety of cancer drugs, including methotrexate, which is used to treat multiple cancer types. Variants within these genes are known to cause changes in drug efficacy via altered levels of drug bioavailability and can increase risk of drug toxicity when drugs are not cleared from the system in time.

4.2.1 ATP-Binding Cassette sub-family B member 1 - *ABCB1*

The *ABCB1* gene has been well characterised for association of specific variants in adverse reactions with multiple drug types. This drug transporter has been associated with both toxicity and efficacy effects, depending on the type of drug and the disease being treated (Hodges *et al.*, 2012).

In my study, only one *ABCB1* LoF variant, a novel variant at position chr7:87195405, was detected in the heterozygous state in a single Ethiopian individual. Since low coverage sequencing data is not ideal for calling rare variants, this LoF variant needs to be validated. To date no *ABCB1* LoF have been described. Of the 28 missense variants in *ABCB1*, 10 were predicted to be likely deleterious to gene function using the annotation tools in this study. All 10 are rare, and none have any record of prior clinical annotation or experimental functional studies. rs148718120 is present only in Gambians in four individuals in the heterozygous state. All variants from this set are extremely rare. From the PharmGKB database, there are five additional coding variants with clinical annotations that occur at much higher frequencies. Two synonymous variants, rs1045642 and rs1128503, have been extremely well characterised clinically, even though they are likely non-functional based on the fact that they do not change the amino acid sequence of the protein.

rs1045642 (also referred to as the *ABCB1* C3435T mutant) has been linked to toxicity effects for methotrexate (a widely used chemotherapy). It has five level 2 annotations from PharmGKB for methotrexate, nevirapine, ondansetron (anti-nausea), digoxin (cardiovascular flutter), and anaesthetics. Key literature reports for methotrexate and nevirapine are summarised in Table 4.1.

Table 4. 1 Key findings from literature for the high impact variant: rs1045642

Drug	Alleles	Effect	Sample Size	Population	p-value	Reference
Methotrexate	AA + AG	Toxicity	110	Lebanese	N/A	(Zgheib *et al.*, 2014)
	AA	Toxicity	332	Asian	0.004	(Suthandiram *et al.*, 2014)
	AA	Toxicity	552	Danish	$p < 0.05$	(Gregers *et al.*, 2015)
	G + A	No effect	332	Chinese	$p > 0.05$	(Liu *et al.*, 2017)
Nevirapine	AA	Toxicity	158	Mozambican	0.038	(Ciccacci *et al.*, 2010)
	AA	Toxicity	97	Mixed	0.021	(Ritchie *et al.*, 2006)
	AA	Toxicity	164	Mixed	0.016	(Haas *et al.*, 2006)

Contradicting reports exist regarding methotrexate treatments, with most studies presenting toxic effects linked to the AA or AG genotypes, however another study found neither allele linked to toxic reactions (Liu *et al.*, 2017). Although there is no available literature regarding methotrexate treatment in Africans, other pharmaceuticals have been reported for toxic effects linked to this variant in Africans. There have been reports for nevirapine (an antiretroviral drug) toxicity associated with the homozygous A allele state in Mozambicans. The mixed participant cohorts studied via Ritchie *et al.* 2006 and Haas *et al.* 2006 consisted of 18% African Americans and 75% Black South Africans respectively. These findings suggest that this variant is functionally relevant in African populations and motivates further clinical research studies for cancer treatments.

The synonymous rs1045642 variant is most likely linked with another functional variant or variant series that could be the cause of the adverse drug reactions associated with it. It is an extremely common variant, found across all the continental African populations assessed with frequencies ranging from 10 to 19%. This variant is far more common in other global populations (>40%), but its frequency in Africans is high enough to warrant further investigation. Due to its distribution across different African populations, as well as evidence of effects noted with other pharmaceuticals in African and African American populations, additional research of rs1045642 is necessary in African individuals currently undergoing cancer treatment. Whole genome or whole exome sequencing would be useful to characterise the entire gene region, especially regarding linkage disequilibrium to assess if rs1045642 is linked to other variants which may contribute to adverse reactions.

The rs1128503 variant has clinical annotations for efficacy of a combination of cancer drugs (Caronia *et al.*, 2011), including methotrexate. Caronia *et al.* found the AG and GG genotypes to be linked to a 50% decrease in expected survival time after treatment. This study also found this variant to be in complete linkage disequilibrium with another variant; rs10276036 (an intronic variant). A small study (n=18) found this variant to be linked to overall toxicity (p= 0.013) following methotrexate treatment in African Americans being treated for rheumatoid arthritis (Ranganathan *et al.*, 2008). rs1128503 also has been reported for toxicity of hormonal- and cell signalling based drugs such as gefitinib (Ma *et al.*, 2017), an epidermal growth factor receptor inhibitor commonly used in breast cancer treatment, and imatinib (Harivenkatesh *et al.*, 2017) and sunitinib (Chu *et al.*, 2015), which target tyrosine kinase systems. These reports are for Asian populations, and are contradicted by other studies that assessed gefitinib (Kobayashi *et al.*, 2015), imatinib (Ben Hassine *et al.*, 2017), and sunitinib (Kim *et al.*, 2013).

The frequencies for the rs1128503 variant allele in continental Africans ranges from ~10 to16%, but this variant lacks clinical information, as well as consensus on its functional effect. However, the small study indicating its role in methotrexate in African Americans encourages more work on the African continent, with larger sample numbers.

One intronic variant, rs2032583, has been noted to be involved with efficacy and likelihood of remission during treatment with antidepressants in Europeans (de Klerk *et al.*, 2013), but no reports are available for cancer treatments. This variant is uncommon in all African populations (<2%), lacks in-depth clinical reports, and is not likely functional given by low CADD and FATHMM scores, thus further study is not recommended.

Of the *ABCB1*missense variants that were originally removed by filtering in this study, the most interesting are rs2229109 and rs2032582, which are both missense variants with well characterised clinical reports. The rs2229109 variant TT and CT genotypes have been linked to an increased risk of diarrhoea and vomiting during treatment with R-CHOP type regimens (a combination of drugs including cyclophosphamide and vincristine used to treat non-Hodgkin's lymphoma) in a cohort of populations of unknown ethnicity (n=760, p=0.04) (Jordheim *et al.*, 2015). It has also been associated with an increased chance of remission and toxicity in a mixed population cohort of children being treated with methotrexate, doxorubicin, vincristine and other drugs (n=164, p=0.0001) (Gregers *et al.*, 2015). This variant was scored low and likely benign/tolerated by SIFT and Polyphen-2, due to a similar change of amino acid from a serine (neutral) to asparagine (neutral), whereas FATHMM-MKL indicated a low confidence positive prediction, and a CADD score of 14.05, which is far below that of other missense variants in this gene. The A allele of this variant is rare in Africans, present only in the Gambian, Luhya (Kenyan) and Ethiopian populations (0.4%, 0.6% and 0.2% respectively). Despite its rarity, strong evidence of deleterious effect on gene function from clinical reports encourages further study in African populations.

The rs2032582 variant was more surprising, as it missed detection due to SIFT, Polyhen-2 failing to produce a score for this variant (which may be due to reference effects, or tri-allelic status, or a combination of these) and FATHMM-MKL scoring low, with only CADD reporting a high likelihood score of 23.2. This variant has two Level 2 annotations, for efficacy of ondansetron and simvastatin (statin). Key findings from literature are summarised for its role in cancer treatments in Table 4.2 below.

Table 4. 2 Key findings from literature for the high impact variant: rs2032582

Drug	Alleles	Effect	Sample Size	Population	p-value	Reference
Sunitinib	AA	Poor Response	90	Asian	0.01	(Chu *et al.*, 2015)
Cytarabine and anthracyclines	CC	Increased Response	101	Asian	0.04	(Kim *et al.*, 2006)
	CC	Decreased Response	100	European	0.02	(Green *et al.*, 2012)
Platinum Compounds	CC	Decreased Toxicity	188	Asian	0.04	(Kim *et al.*, 2009)

The rs2032582 variant has been linked both to improved and decreased efficacy of a variety of cancer treatments. While the AA genotype may improve response to sunitinib, the role of the CC genotype is more difficult to understand, due to conflicting findings. The same CC genotype was found to have effects on both anthracyclines and platinum compounds, indicating that this variant can have varying effects on different drug types. As the C allele is common in Africans, most patients with this variant may be advised to avoid platinum-based treatment due a higher likelihood for toxicity.

The minor allele A is rare in African populations compared to the other global population groups. Frequencies of the A allele are distinct among African populations, with Mende, Luhya and Zulu populations (~1.6%) having substantially greater frequencies than Gambians, and Bagandans (<0.5% on average), and with an absence in both Nigerian populations. Because of the involvement of multiple drugs, for both toxicity and efficacy, as well as high functional likelihood indicated by CADD, the rs2032582 variant should be further investigated in Africans, particularly in Mende, Luyha and Zulu populations.

4.2.2 Conclusion: *ABCB1* variants

The *ABCB1* gene contains African population specific rare variants with high likelihood of deleterious effects (annotated in this study), as well as known variants of clinical impact whose frequencies are different from other global populations. This gene should be examined in detail within African cancer patients, most particularly in those who are having adverse reactions to methotrexate, imatinib, or sunitinib type regimens, or patients who still present with adverse

effects despite different drug regiment types used for subsequent treatments. These studies would need reasonable sample sizes with at least 100 patients (as seen in other study cohorts), deep sequencing on a population level to better characterise variants such as rs2032582 and rs2229109, as well as laboratory validation to verify their functional effects. This knowledge could be used to design preventative guidelines to avoid adverse reactions regarding in African patients.

4.2.3 Solute Carrier Family 19 Member 1 - *SLC19A1*

The *SLC19A1* gene encodes RFC1, which is key to folate transport system protein. Folate is needed to maintain cell growth and DNA synthesis. Drugs such as methotrexate target the folate transport system in order to slow the growth of cancer, however, variants in the *SLC19A1* gene are known to affect treatment response. Two variants were predicted to have damaging effects on gene function, rs202074305 and rs557998866, both of which are extremely rare (<0.5%) and limited to specific populations. rs202074305 was present only in Yoruba and rs557998866 present only in Bagandan and Ethiopians populations. Neither variant has any clinical annotations, which may be due to their rarity. One variant, rs1051266, was identified in the PharmGKB datasets for potential clinical effects.

The rs1051266 missense variant had very low scores for all predictive tools and so was not included following filtering. However, it has multiple clinical reports linking it to adverse drug reactions, but these reports present conflicting findings. The bioinformatics tools predict this variant to be likely benign, as the amino acid change does not appear to cause large structural changes. The variant allele is also common worldwide (the T allele is present in high frequencies globally, with an average of ~50%). The clinical relevance of this variant has been difficult to characterise, due to numerous conflicting findings. The main findings are related to methotrexate, with results for both toxicity as well as efficacy of the drug. Toxicity resulting from methotrexate treatment has been found related to the TT homozygous genotype in a cohort of Spanish children with lymphoblastic leukemia (n=141, p=0.008) (Salazar *et al.*, 2012); a small Asian cohort with lymphoblastic leukemia (n=13, p=0.035) (Imanishi *et al.*, 2007) and a case/control Asian cohort with lymphoblastic leukemia or non-Hodgkin lymphoma (n=87, 51 cases, 0.032) (Suthandiram *et al.*, 2014). Other sources found no association with toxicity for the TT genotype compared to CC or CT in a cohort of Asian patients (n=135, p=0.93) (He *et al.*, 2014) and in a large cohort of Chinese children with lymphoblastic leukemia (n=322, p=0.384) (Liu *et al.*, 2017). This means that the predictive tools may be correct in their prediction that this variant may not be inducing a strong functional effect, however, these

effects may not be limited to genetic factors involving *SLC19A1* alone, but also could involve environmental factors or could be caused by other variants or genes. Further work is needed to clearly define allelic effects with regard to drug response. Though controversial evidence exists, this variant should be further researched in African populations, where the T allele is common (~70%), particularly in patients undergoing treatment with drugs targeting folate transport.

The *SLC19A1* gene should be investigated in a similar manner to those motivated for *ABCB1*, particularly with deep sequencing and laboratory validation, which could help characterise variants such as rs202074305 and rs557998866 and clarify the findings of the rs1051266 variant.

4.2.4 Conclusions: Drug transporters

Two key drug transporter genes, *ABCB1* and *SLC19A1*, were investigated and were found to have variants of predicted functional impact, and others which have known clinical evidence but that require replication studies in continental Africans. *ABCB1* variants could affect numerous drug types, that include treatment for a large variety of cancer types, and *SLC19A1* variants could affect methotrexate and other folate system targeting drugs, which are relevant in a variety of cancer types. These broad use treatments will be more relevant as cancer rates continue to rise in Africa, and key PGx knowledge could help prevent adverse reactions, and improve outcomes for African cancer patients.

4.3 Key drug targets – *CYP19A1* and *ESR1*

Two genes encoding key drug targets for commonly used cancer drugs, *CYP19A1*, and *ESR1*, were investigated. Changes in their structure or expression can result in reduced treatment efficacy, causing lower chances of survival and higher chances of cancer recurrence.

4.3.1 Human Aromatase - *CYP19A1*

CYP19A1, which encodes the human aromatase enzyme, is a key target of aromatase inhibitor type drugs. Although it is a cytochrome P450 enzyme, its role regarding PGx significance is most well characterised as a drug target.

The *CYP19A1* gene had only two missense variants predicted to be deleterious and both were rare and limited to a single population: rs142652579 in the Esan, and rs143562020 in the Zulu. Both had no prior clinical annotations but are predicted as likely deleterious with high scores for all toolsets. Validation of these variants would be required to assess their PGx significance in Africans.

Three non-coding variants, rs1008805, rs4646, rs6493497, were found in the PharmGKB dataset with clinical relevance. rs1008805 is an intronic variant, found associated with joint pain during treatment with anastrozole, an aromatase inhibitor commonly used to treat breast cancer, in a European cohort of postmenopausal breast cancer patients (n=104, p=0.004) (Gervasini *et al.*, 2017a). This variant is common across the African populations analysed but has no clinical or functional reports in Africans.

The rs4646 variant, located in the 3' UTR, is the most well characterised *CYP19A1* variant to date. The AA and AC genotypes were associated with better response to numerous chemotherapies in a cohort of Asian women with hormone receptor positive breast cancer (n=294, p=0.011) (Shao *et al.*, 2015a) as well as with better response to tamoxifen treatment in a similar cohort of early breast cancer patients (Shao *et al.*, 2015b). This is interesting to note as the A allele is the minor allele globally and is found to improve treatment outcome compared to the more common CC homozygous genotype. African A allele frequencies range from 28 to 33%, which is important to note regarding prognoses for African breast cancer patients.

An upstream variant (rs6493497) has been found associated with increased aromatase activity following treatment with anastrozole, exemestane and letrozole (aromatase inhibitors) which may affect treatment outcome (Wang *et al.*, 2010). This effect is also linked to the minor allele, A, and was noted in heterozygotes in the study, but no homozygous AA genotypes were detected in the cohort. The rs749292 intronic variant, conversely to the others in this gene which primarily affect drug efficacy, has been associated with a toxic effect following exemestane and letrozole treatment for low bone density (is this correct?) in a cohort of mixed populations (n=224, p=4.6E-8) (Oesterreich *et al.*, 2015). Frequencies for both rs6493497 (~33%) and rs749292 (~34-49%) in Africans are very high, and thus motivate further study of these variants in African patients with hormone receptor positive breast cancer.

4.3.2 Oestrogen Receptor 1 - *ESR1*

The *ESR1* gene, linked most closely to breast cancer treatments, had only one missense variant, rs188957694, predicted to be likely deleterious. Although the predictions were in consensus for all toolsets, no clinical annotations are available. This may be due to it being extremely rare, present in only one heterozygote in the Yoruba population and in no other global populations within the KGP.

Interestingly, a very rare *ESR1* variant, with potential LoF effect, COSM325271, was not included as functionally significant following the scoring method. It was originally annotated as a start lost variant, but was classified as low confidence by the LOFTEE tool, which was later confirmed via the COSMIC analysis data, which found this variant to be a start retained variant (Rudin *et al.*, 2012). Meaning it is a variant where a change occurs within the start site, but transcription start is retained. This variant is also extremely rare, present only in the Ethiopian population, and elsewhere only in one sample from the COSMIC study. Without further confirmation for the existence of this variant (e.g. using deep sequencing or Sanger sequencing to confirm that it is not an artefact) and functional assays to assess its impact on the gene, it is not likely to be useful for pharmacogenomic use in Africa.

Two additional intronic variants were included due to clinical evidence related to adverse reactions. The rs9340799 and rs2234693 variants have been found linked to adverse reactions in oestrogen-based drugs and drugs targeting oestrogen pathways. These variants are located very close together, and were found in linkage disequilibrium with each other in studies linking them to adverse reactions involving anastrozole and letrozole treatment (Wang *et al.*, 2013). This study found the rs9340799 A allele and the rs2234693 C allele to be associated with increased musculoskeletal pain, arthritis, and diminished joint function, following the use of these aromatase inhibitors, in a case control cohort of Asian patients with early breast cancer (n=436, 206 cases, $p=9.49x10^{-7}$). These alleles are common in Africans, with the rs9340799 A allele present at ~70% and the rs2234693 C allele present at ~60% average frequencies in Africans. The rs9340799 A allele is present in all other global populations at similar frequencies to those of Africans, but is more common in East Asian populations (81%). The rs2234693 C allele is also common in other global populations, but in significantly lower frequencies than Africans (~39%).

Although these variants are also common in European populations, no adverse reactions have been reported linked to these variants. This suggests that there may be additional factors, potentially the actual causal variants, that co-segregate with these intronic variants in Asian populations. Intronic variants may play a role in gene expression, which often includes numerous additional factors. Although these variants are very common in Africans populations, limited evidence (which is specific to Asian populations) places doubt onto the true functional relevance of the rs2234693 and rs9340799 variants. Further research into African populations would not be well motivated until future evidence regarding function is present.

4.3.3 Conclusions: Key Drug Targets

The *CYP19A1* and *ESR1* genes play key roles with aromatase inhibitors used to treat breast cancer, and with breast cancer being one of the two most common cancer types in females in Africa, should not be omitted from trials investigating adverse reactions with these drug types. Few missense variants were detected as potential PGx markers in Africans, all of which are rare and limited to specific populations. Effects of non-coding variants which likely effect gene regulation have been found in Europeans and Asian populations. These variants are common in African populations, so future studies should note that regulatory variant effects on these genes may be more commonly involved in adverse drug reactions than coding variant effects.

4.4 Drug Metabolisers

Drug metabolising enzymes are of critical relevance to PGx knowledge which aims to ensure drug safety. These enzymes are generally the rate limiting factor of pathways to break down drugs into by-products which can be safely eliminated from the body. A key aim of PGx is to identify variants that alter gene function, in terms of affecting the activity levels of the respective enzyme. The CPIC has aimed to standardise terms to report the activity level of CYP P450 enzymes to describe the metaboliser status of a patient (Caudle *et al.*, 2017). These guidelines stratify patients into ultra-rapid, rapid and normal metabolisers, which correspond to normal or increased enzyme activity, and into intermediate and poor metabolisers, with reduced enzyme activity compared to individuals with normal levels. Variants that alter protein structure have a large impact to activity, and these impacts differ between cases of heterozygous or homozygous genotypes such that two deleterious alleles have a larger impact on activity than of one deleterious and one normal functioning allele.

Cytochrome P450 genes encode for a family of drug metabolising enzymes which are responsible for the metabolism of 90% of drugs currently in use. Characterising variants that can alter the function of these genes allows for the development of strategies to guide treatment to avert adverse reactions and prevent toxicity.

4.4.1 Cytochrome P450 Family 1 Subfamily B Member 1 - *CYP1B1*

CYP1B1, a member of the cytochrome P450 family, has 13 missense variants across the African populations assessed, with only three missense variants (rs1800440, rs747324108, rs9282671) classified as likely functionally deleterious by the bioinformatics tools in this study, and one additional missense variant (rs1056836), classified as likely non-functional, but which has significant prior clinical annotations.

The rs1800440 variant, has been cited for not affecting treatment regimens including capecitabine, cisplatin, and docetaxel in a cohort of Europeans with pancreatic neoplasms (n=121, p=0.8) (Giovannetti *et al.*, 2011). Two others, rs747324108, which is African specific, and rs9282671, which is rare globally, have no clinical evidence available.

The rs1056836 variant, classified as likely benign by all toolsets, (applied to the G allele) induces a change from valine to leucine, both very similar, small non-polar amino acids. The G allele is far more common in non-Africans, with frequencies of 60% in Europeans and 83-93% in Asian populations, while African frequencies range between 12-20%, except for Ethiopians in whom the frequency is 40%. There are conflicting results for this variant with regard to effect, with some studies reporting the CC genotype associated with poor response to cyclophosphamide, epirubicin, and fluorouracil (Le Morvan *et al.*, 2015) and decreased overall survival following treatment with docetaxel (Sissung *et al.*, 2008); while others report no association with the same drug or similar drug types (Giovannetti *et al.*, 2011; Hertz *et al.*, 2012). Although the reports are conflicting, the C allele is very common in African populations, urging further study to ascertain its role, if any, in adverse drug reactions in Africans.

4.4.2 Cytochrome P450 Family 2 Subfamily D Member 6 - *CYP2D6*

CYP2D6 had the greatest number of missense variants of all the CYP450 gene family members assessed here, with 24 across all populations, as well as 3 LoF variants, one of which has prior clinical annotations.

Of the LoF variants detected in this gene, only rs3892097, a splice acceptor variant, is common to all Africans, and has prior clinical annotations, whereas the other two (rs147960066 and rs536109057), are both stop gain variants which are rare, with each present only in one population. The stop gains have no prior reports of any confirmed functional effect. The rs3892097 (commonly referred to as *CYP2D6*4*) variant has been well characterised in terms of function, as it is used for dosing indications for numerous antidepressants (Hicks *et al.*, 2015). This variant is most involved with adverse reactions in treatment with tamoxifen, a widely used drug used to prevent breast cancer recurrence in hormone receptor positive breast cancers. The TT genotype has been associated with an increased risk of cancer relapse in a clinical trial of North American women treated with tamoxifen (n=190, p=0.02) (Goetz *et al.*, 2005); a subset of Austrian patients treated with anthracyclines and tamoxifen (n=144, p=0.034) (Stingl *et al.*, 2010); and additional studies on the CT heterozygous genotypes (along with TT) in a large cohort of German patients (n=1325, p=007) (Schroth *et al.*, 2009). The T

allele is most common in Europeans (19%), and less common in Africans (~8%). No reports exist for tamoxifen and *CYP2D6* function in Africans. With breast cancers rates increasing rapidly in Africa, further study of this variant in clinical trials, along with genotyping of this variant prior to treatment is advised.

Of 24 missense variants, nine were predicted likely deleterious, with only one of these (rs1065852) having prior records of clinical effect. One additional missense variant, rs16947, was found to have clinical annotations but had low predictive scores for all toolsets. The rs1065852 A allele has been associated with an increased risk of breast cancer recurrence during tamoxifen treatment, in a cohort of German patients, for whom their ethnicities were not stated (n=206, p=0.02) (Schroth *et al.*, 2007). Adverse reactions linked to antipsychotics have also been associated with the AA genotype in both European (Ji *et al.*, 2014) and Asian populations (Han *et al.*, 2013). This variant is present in all African populations, but at different frequency ranges for different African regions. The frequency of the A allele in West Africans ranges from 9-16%, whereas that in East Africans it ranges from 3.5-7%. The South African Zulu have a frequency of 11%. This variant may disproportionately affect West and South Africans more than East Africans in terms of adverse reactions following tamoxifen treatment. Internationally, this variant is rarer in Africans compared to other populations, with the A allele present at frequencies of 20% in Europeans and 57% of East Asians. Further work is recommended to characterise this variant in African populations, with consideration that it may more likely be involved in adverse drug reactions in West African individuals.

The rs16947 variant genotypes AA and AG have been associated with an increased risk of brachycardia (very low heart rate) in Asian glaucoma patients treated with timolol (a beta blocker used to lower inner eye pressure) (n=123, p=0.043) (Yuan *et al.*, 2010). The A allele has also been found to decrease the expression levels of the *CYP2D6* gene in cell culture of human liver cells (Wang *et al.*, 2014). The amino acid change in this case is a switch from cysteine (polar) to arginine (charged) and it scored low based on the type of amino acid change and position (SIFT and Polyphen-2 scores indicating tolerated and benign), and with low scores with respect to conservation based on multiple sequence alignments and ancestry (FATHMM and CADD). The variant is not particularly conserved worldwide, with the G allele (with A as the minor allele) present at frequencies of 67% in Admixed-American, 66% of Europeans, 64% in South Asians and 86% in East Asians. Conversely, the A allele is the major allele in African populations compare to G, being present at frequencies of almost 60%, which is significantly different from all other global populations. Despite a lack of evidence-based association with

CYP2D6 function particularly for interactions with cancer drugs, further studies should be conducted in Africans due to the frequencies being much higher in Africans for the alleles found associated with other drug types, as well as with *CYP2D6* expression levels.

Of the eight missense variants with high scoring predictions, but no clinical reports available, the rs532668079 (Esan), rs376636053 (Mende), rs373243894 (Luhya), rs143145507 (Bagandan) variants are all found in only one population, and are rare, with frequencies of ≤1%, and thus may not be useful for population pharmacogenomics studies. None of these variants is significantly different across the African populations, due to their rarity, although this may be due to the small sample sizes of each population. The rs1058172, rs28371703, and rs369177208 variants are also rare, but occur in multiple populations. The rs1058172 and rs28371703 variants, show significantly different allele frequencies between Africans, with the rs1058172 T allele present in Zulu (0.5%) and Ethiopians (3%); and rs28371703 T allele present in all but the Mende and Bagandan populations, at an allele frequency range of 0.5-3%, and with Ethiopians having the highest frequency (3%). These two variants are likely damaging to gene function, with all annotation tools in consensus for rs1058172, and for rs28371703 which only scored low by FATHMM-MKL. With rs28371703 being rarer in Africans than other global populations, it would be expected that other studies in well characterised populations would have already identified this variant regarding function. More doubt on its functional relevance is presented by a study, although dated, which found this variant to not affect *CYP2D6* activity on bufuralol (beta blocker) metabolism within a kidney cell line expression system (Kagimoto *et al.*, 1990). The rs1059172 variant was not present in any other populations within the KGP (this variant was only present in the AGVP population dataset), but has since been found in multiple populations within the ExAC project aggregation datasets (Lek *et al.*, 2016), at frequencies of 14% for non-Finnish Europeans, 11% of Ashkenazi Jewish, and 2% of the African samples within these datasets. This places the rs1059172 variant in a similar category as that of rs28371703, where there are strong predictions, but no clinical reports found for this variant.

The rs59421388 missense variant is a particularly interesting variant. It was annotated with high confidence of likely deleterious effect by all toolsets, and is a change from valine (small and non-polar) to methionine (polar), which is predicted to be deleterious and damaging to protein function by SIFT and Polyphen-2 respectively. This variant is also African specific, occurring in all African populations, but at a significantly varying frequency across African populations. It is most common in the Bagandan and Luyha populations, at frequencies of 22%

and 17% respectively, with the West African Esan and Yoruba having lower frequencies at ~10%, and rarest in Ethiopians, at 2%. This variant is present in the *CYP2D6*29* haplotype along with rs16947 (as described above), rs61736512 and rs1135840 (not included in this study). This haplotype is common to 20% of KGP Africans including the African diaspora populations (haplotype frequency retrieved from Ensembl). This haplotype has been assessed before in Tanzanians receiving debrisoquine, and was found to significantly reduce enzyme activity with this drug (Wennerholm *et al.*, 2001). No prior clinical information or annotations are present for this variant for cancer drugs as it is only present in African populations, urging further work on the African continent to characterise this variant in clinical trials, particularly for tamoxifen-based treatments, as well as for psychiatric drugs and beta blockers which are metabolized by *CYP2D6*.

The *CYP2D6* gene is also known to affect drug reactions due to CNVs of the entire gene sequence, with increasing gene copy numbers producing an ultra-rapid metaboliser phenotype (Ingelman-Sundberg, 2005). CNVs were not assessed in this study but can play key roles in drug response. Duplications of the *CYP2D6* gene can result in drugs being metabolised too rapidly and losing their therapeutic effect. *CYP2D6* CNVs of more than 3 copies are present in 9% of African Americans, 4.6% of Caucasian Americans, and 6.2% of Asians as assessed in a large study of CNVs in the USA (Beoris *et al.*, 2016). Considerations should be taken in studies of continental Africans that while SNVs may not be present, CNVs may also induce adverse drug reactions, and will require more specialised assays than standard genotyping (Langaee *et al.*, 2015).

4.4.3 Cytochrome P450 Family 3 Subfamily A Member 4 - *CYP3A4*

CYP3A4 had no detected LoF variants, only a set of 13 missense variants, of which seven African specific variants were predicted to be deleterious, with only two variants with consensus from all toolsets: a novel variant, 7:99375691 (present only in the Zulu population), and rs897005504 (present in Zulu and Ethiopian populations). Both are rare, with allele frequencies of 0.5%. The other variants are also rare, with most being limited to one population, except for the rs12721629 and rs57409622 variants. rs12721629 is present in the Luhya (2%), Zulu (2%) and Bagandan populations (1%), and more rarely in the Esan population (0.5%). This variant is significantly different among African populations and is more likely to be present in East and South Africans. The rs57409622 variant is present in the West African Esan, Gambian and Mende at ~0.5% frequency. These variants have not yet been characterised, but their high scoring functional predictions motivate further study. These studies should be

directed in the case of rs12721629, which is more likely to be found in East Africans and South African Zulu populations.

There are four additional variants with clinical annotations that were not detected in this study due to being intronic or upstream region variants. The rs2740574 variant (also referred to as *CYP3A4*1B*) has been well characterised in terms of associations with numerous cancer drugs and has one Level 2 annotation for tacrolimus (an immunosuppressant). Key findings from the literature for these are displayed in Table 4.3 below. This upstream promotor variant was scored as low likelihood of effect by both CADD and FATHMM, which may have missed this variant due to its higher frequencies (as these tools use sequence conservation as part of their scoring). The T allele is the common allele globally, with an overall frequency of 77% in the KGP global populations, but it is the minor allele in Africans, with an average frequency of 26% in the African populations analysed here.

Table 4. 3 Key findings from literature for the high impact variant: rs2740574

Drug	Alleles	Effect	Sample Size	Population	p-value	Reference
Tacrolimus	CT	Changed Dosage	206	European	p<0.01	(Tavira *et al.*, 2013)
	CT	Changed Dosage	446	Mixed Populations	$6.4x10^{-9}$	(Birdwell *et al.*, 2012)
	CT	Changed Dosage	64	Mixed Populations	0.027	(Hesselink *et al.*, 2003)
	C	No effect	298	Mixed Populations	N/A	(de Jonge *et al.*, 2011)
	C	No effect	014	Brazilian	0.79	(Santoro *et al.*, 2013)
Tamoxifen	C	Increased recurrence	126	Canadian	0.004	(Chu *et al.*, 2007)
Taxanes	CC+CT	Toxicity	70	European	0.01	(Boso *et al.*, 2014)
	CC	Toxicity	217	Turkish	0.038	(Kus *et al.*, 2016)
Cyclophosphamide	CC	Reduced Toxicity	78	Mixed Population	0.01	(Su *et al.*, 2010)

For the studies that found the variant allele to be linked to dosage change, all patients required higher doses of the drug to maintain effective treatment. There are two studies however, that found no change in dose necessary for patients with the same genotypes. Other reports indicate that the genotype of the kidney donor, not of the patient, has an effect (Gervasini *et al.*, 2017b). The relationship between this variant and tacrolimus may be an indication of possible effects with other drug types, including cancer treatments, where dosages may be more difficult to adjust. There is also an indirect effect on renal cancer patients, who have kidney transplants as part of their treatment. It would be relevant to note that the genotypes of both patient and kidney donor are important in these cases. The rs2740574 CT genotype and C allele have been linked to toxicity effects related to key cancer treatments often used for breast cancer. It is critical to

note that should these drugs be used in Africans, the C allele is far more common across all African populations, with frequencies of ~83% in Mende and Luhya populations, 80-75% in Yoruba, Esan, Zulu and Bagandans, and 52% of Ethiopians. Not all effects reported for cancer drugs are negative for this allele, as the CC genotype has been found associated with longer times to chemotherapy induced ovarian failure following treatment with cyclophosphamide (Su *et al.*, 2010). With the C allele highly common in Africans (with the exception of Ethiopians), and major effects (both positive and negative) noted for multiple drug types, further research of this variant is urged, particularly in African breast cancer patients.

The three other *CYP3A4* variants, rs2242480, rs35599367, and rs4646440, are all intronic variants, but have associations with adverse reactions for methadone (an opioid painkiller, often used to treat addiction withdrawal). The rs2242480 variant, like rs2740574, has the minor and major alleles inverted for Africans compared to other global populations, with C being the major allele globally, but the minor allele at ~15% average frequency in Africans. The CT and TT genotypes have been associated with methadone toxicity in an Asian cohort (n=732, p=0.0085) (Chen *et al.*, 2011). The rs4646440 variant has also been linked to adverse reactions with methadone in the same cohort (n=732, p=0.028) (Chen *et al.*, 2011), but is varied in frequency across African populations, being more common in the Bagandan, Luhya, Esan, Yoruba and Zulu populations (~12%), and rarer in Gambian and Mende (~4%). No reports are available for any cancer treatment associated with rs2242480 and rs4646440, but due the high frequency of the rs2242480 C allele in Africans, this variant should be analysed in more detail in Africans, and it should be recommended that other painkillers than methadone be used in African populations. The rs35599367 variant, is the rarest of all variants found to have clinical annotations, present only in Ethiopians at (1%), and in Europeans (5%) and Admixed-Americans (3%). The AG genotype has been reported for recommendation for increased everolimus (chemotherapy) in a small European cohort (n=37, p=0.019) (Pascual *et al.*, 2017), but as it is uncommon in Africans, and no data exists related to drug toxicity, further investigation in Africans is not recommended.

4.4.4 Cytochrome P450 Family 3 Subfamily A Member 5 - *CYP3A5*

Variant in *CYP3A5* are often found to be related to adverse drug reactions and along with *CYP3A4*, has the greatest number of LoF variants detected in this study. The four LoF variants include three rare (with frequencies of <1%) and population specific variants: two stop gains, rs149664815 and rs111371159, and the rs200579169 frameshift variant. The rs149664815 was present in the Ethiopian and the Luyha populations, the rs200579169 variant was only present

in the Ethiopian population, and the rs111371159 variant was present only in the Gambian population. These variants have no clinical annotations or reports, but this may be due to their rarity, with only the rs200579169 frameshift variant occurring in another global population, present in Europeans at ~1% frequency. Due to these variants having a complete knockdown of gene function, they should be researched in more detail, however, their rarity, population specificity and lack of clinical reports may lead these to have low significance for pharmacogenomics in Africa.

The fourth LoF variant, the rs4130334 frameshift variant (*CYP3A5*7*), is present in all African populations, and is African-specific, present only in Africans and African diaspora populations. The frameshift is induced via an insertion of adenine (henceforth referred to as A), which forms the genotypes of A/del, A/A and del/del, where the deletion in this case is the absence of the insertion. The frequencies of the A allele are varied between the African populations, with the allele being significantly rarer in Ethiopians (<1%) than in the other populations (~11%). This variant does have clinical reports for adverse drug reactions, mostly for tacrolimus and lumefantrine (an antimalarial drug). The A allele is associated with decreased clearance of tacrolimus in a Brazilian cohort of kidney transplant recipients (n=151, p=1.0E-4) (Santoro *et al.*, 2011), requiring these patients to have lower drug dosages to avoid toxicity. The Brazilian population has a substantial portion of African admixture (Kehdy *et al.*, 2015), which explains the presence of this African variant. The A allele has also been shown to affect tacrolimus clearance in a small cohort (unknown population group), noting that the rate of drug clearance is affected in heterozygote A/del genotypes, as well as in A/A genotypes (n=24, p=3.0E-4) (Zheng *et al.*, 2012). These findings are supported in an analysis of these genotypes compared to other common *CYP3A5* allele genotypes, in a mixed ethnicity cohort of kidney transplant recipients (n=162, p=0.029) (Maldonado *et al.*, 2017). This variant, like those in *CYP3A4*, may affect renal cancer patients who have had kidney transplants. Further research is urged in Africans, particularly in those receiving kidney transplants, as well in cases where an African individual may be the donor, as *CYP3A5* variants have been noted in donor kidney response as well (Gervasini *et al.*, 2017b).

Additional evidence supporting the effect of the rs4130334 variant in Africans was noted in Tanzanian malaria patients treated with lumefantrine, of which the A/A and A/del genotypes were found associated with increased drug concentration levels compared to patients with del/del genotypes (n=92, p=0.039) (Mutagonda *et al.*, 2017).

Regarding LoF variants in the *CYP3A5* gene, the only LoF variant with additional clinical data was rs776746, named an intronic variant in the canonical transcript of *CYP3A5* as defined by VEP. This variant was defined as low confidence by the LOFTEE tool possibly due to it being present in a shorter transcript of the gene and not strongly conserved, but as it has substantial clinical reports available, the variant was included in this analysis. This variant is the most common *CYP3A5* mutation, which has been named the *CYP3A5*3* allele, it is a splice acceptor variant, but is present in the less common, second longest transcript of *CYP3A5*. This presents a situation where the canonical transcript as defined by Ensembl – usually the longest transcript and most well characterised - is not always the most relevant transcript in terms of gene expression in humans. The C allele induces the splice error, which results in a retained intron and a non-functional transcript. The T allele, is the minor allele in all other global populations (versus C), but is the major allele in Africans, present at ~83% frequency in all African populations assessed except Ethiopians, with frequencies of 50%, indicating that most Africans would have functional alleles, and thus are more likely to exhibit normal metabolism.

The rs776746 variant, like others in *CYP3A4* and *CYP3A5*, has been well characterised with regard to tacrolimus dosage (with a Level 1 annotation) in Asian (Niioka *et al.*, 2012), European (Tavira *et al.*, 2013), and African American (Jacobson *et al.*, 2011) cohorts. The expressers of functional CYP3A5, will require an increased dose to prevent transplant rejection, but non-expressers will require lower doses to prevent toxicity of the drug. This means that Africans, based on the rs776746 T allele frequency, should in general require higher doses of the drug compared to other populations, with the exception of Ethiopians. With 50% frequency of expressors and non-expressors, individuals would benefit best from genotyping prior to treatment. Regarding cancer specific treatments, the *CYP3A5* gene variants can cause adverse reactions in an unusual manner. Key findings of these cases are displayed in Table 4.4.

Table 4. 4 Key findings from literature for the high impact variant: rs776746

Drug	Alleles	Effect	Sample Size	Population	p-value	Reference
Paclitaxel	T	Toxicity	118	European	0.012	(Leskela *et al.*, 2011)
Sunitinib	T	Dose Reduction	84	Mixed Population	0.022	(Garcia-Donas *et al.*, 2011)
Imatinib	C	Concentration Increased	110	Nigerian	0.0001	(Adeagbo *et al.*, 2016)

The effect allele of rs776746 (T), has been associated with toxicity associated with paclitaxel (a broad range chemotherapy) in Europeans and with dose reductions related to toxicity from sunitinib in a mixed population, and it is suggested that the metabolites of such drugs in large concentrations may be inducing the toxic effects. This is unusual as the functional allele in this case is associated with toxicity, and thus, may have large implications for Africans where the functional allele is significantly more common. In these cases, the African specific rs4130334 frameshift variant may be protective of toxic effects, and will require studies to verify this effect. The non-functional rs776746 C allele has been found to affect imatinib concentration in a Nigerian cohort, supporting that the variant is present and can actively affect cancer treatments in continental Africans. Studies in continental Africans who are assessing toxicity effects of paclitaxel and sunitinib drugs should strongly consider genotyping of the rs776746 and rs4130334 variants to assess whether *CYP3A5* activity is involved in drug metabolism related adverse reactions.

Missense variation within the *CYP3A5* gene was assessed and five variants of likely deleterious effect were identified with consensus for all tools: two novel variants, 7:99270300 and 7:99270249, and rs142823108, rs145774441, and rs188366390. All of these variants are rare (<1%), and limited to only one population, except for rs188366390, which is present at low frequency in the Zulu, Bagandan and Ethiopian populations. None of these variants have any prior reports for adverse reactions or pharmacokinetic effect. These variants would first require

confirmation via deep sequencing, then be followed up with functional assays to assess their effect on CYP3A5 activity. Due to their rarity, and not being common across Africans or African regions, these variants may not be as useful as assays the LoF variants previously described in the *CYP3A5* gene for use in pharmacogenomic guidance in Africans.

4.4.5 Summary: Cytochrome P450 genes

There are numerous rare missense variants in cytochrome P450 genes that are predicted to be likely deleterious, but have little or no clinical evidence regarding their function. These variants, which are mostly limited to single populations, may not be useful for broad base PGx guidelines, but motivate further work in African populations to better characterise population specific risk factors. There is also the case of the *CYP2D6* rs59421388 variant, which is common in Africans, and predicted likely deleterious with high confidence, but lacks clinical reports for cancer treatments. Clinical studies of this variant and others in the *CYP2D6*29* haplotype group are strongly encouraged for cancer treatments, especially as this haplotype has been found to decrease activity with other drug types. For other missense variants of known clinical relevance, examples such as the *CYP3A4* rs2740574 and the *CYP2D6* rs1065852 variants highlight that there are intra-African population differences in allele frequency which indicate that drug response will not be uniform between populations from different African regions. The diversity of these variants highlights the need to sequence more African samples, and have these data shared across networks to map out African variation for this gene family.

This gene family also presented with the greatest number of LoF variants, which although rare, are more useful for PGx as they can produce more severe side effects via the complete knockout of gene function. These variants are yet uncharacterised clinically, and will require further functional studies to establish their effect. Three of these variants are relatively common and have verified clinical effects. The *CYP3A5* rs4130334 frameshift variant and rs776746 splice acceptor variants, and the *CYP2D6* rs3892097 splice acceptor variant, are common in Africans, and are known to be involved in adverse reactions. Activity assays in African samples are recommended to help characterise treatment models, with consideration that genes such as *CYP3A5* are more likely to be functional in Africans compared to other populations, and will have PGx effects related to function rather than loss.

The diversity of this gene family, in terms of variant characteristics, and frequency variation both within Africa and relative to other global populations, motivates further study of these

genes which should involve sequencing of more populations at deeper coverage, as well as functional assays and clinical validation of possible PGx effects.

4.4.6 Key catalyst – drug metaboliser GSTP1

GSTP1 encodes a broad substrate enzyme that metabolises toxic compounds via conjugation with reduced glutathione. It is well characterised for its role in cancer drug metabolism, and has been linked to adverse reactions to fluoropyrimidine, cyclophosphamides and platinum-based treatments. The *GSTP1* gene contained the fewest variants present within and around its genic region, fewer than all other genes in this analysis. The total number of missense variants was higher than for *TYMS* and *CYP19A1* indicating that *GSTP1*, although having fewer variants, may have higher variability in its coding regions compared to larger genes.

There was only one *GSTP1* missense variant (rs199949339) predicted likely deleterious, although without consensus from all toolsets, with FATHMM-MKL scoring just below the cut-off of significance, which is still indicative of functional impact, but with lower confidence. The rs199949339 variant is rare, present only in the Esan population at a frequency of 1%. This variant is absent from other global populations, and is only found elsewhere in Afro-Caribbean's, who most likely have this variant as a result of their West African ancestry. This variant is not linked to any reported phenotypes, encouraging functional assays to determine its effect, which could prove relevant for the Esan, as well as the West African diaspora globally.

Two additional missense variants, rs1138272 and the rs1695, are included because they have extensive clinical annotation, even though they were excluded by my methodology. The rs1138272 variant has been associated with efficacy of cisplatin treatment but with multiple conflicting reports. The TT and TC genotypes were linked to decreased event-free survival and overall survival in a Slovenian cohort treated with cisplatin (n=66, p=0005) (Goricar *et al.*, 2015). However, the TC genotype was found to have no effect in a Siberian cohort (n=104) (Khrunin *et al.*, 2012), nor in a cohort of unknown population origin (n=108) (Booton *et al.*, 2006). This variant is significantly rarer in Africans (~1%) than in Europeans and South Asian populations (7%). It produced very low scores based on a change between very similar amino acids, alanine and valine, which are both short and non-polar. With this variant being rare in African populations, with inconclusive clinical reports and low functional prediction scores, it is unlikely that this variant will be useful for pharmacogenomic guidelines in Africa.

On the other hand, the rs1695 variant is well characterised with consistent findings in terms of its clinical effect, with multiple Level 2 annotations for cancer drugs (key findings in Table 4.5). The minor allele, G, is common worldwide with an average frequency of 35%, but is more common in Africans (42%) and Admixed Americans (48%). The Ethiopian population has a significantly lower frequency (25%) than the other African populations. The various genotypes of the rs1695 variant have different effects depending on the type of drug used, and key findings of these from literature are displayed in Table 4.5 below.

Table 4. 5 Key findings from literature for the high impact variant: rs1695

Drug	Alleles	Effect	Sample Size	Population	p-value	Reference
Platinum Based drugs	AA	Toxicity	118	Korean	0.015	(Kim *et al.*, 2009)
	AA	Toxicity (with fluorouracil and irinotecan)	299	North American	0.0006	(McLeod *et al.*, 2010)
	AA	Toxicity	106	Mixed Population	0.019	(Stoehlmacher *et al.*, 2004)
	AG + GG	Toxicity (hearing loss)	64	North American	0.03	(Rednam *et al.*, 2013)
	AG+GG	Toxicity (hearing loss)	188	Canadian		(Drogemoller *et al.*, 2017)
Cyclophosphamide and Epirubicin	AA	Increased Response	24	Unknown	0.376	(Oliveira *et al.*, 2010)
	AA	Increased Response	118	Asian	0.024	(Zhang *et al.*, 2011)
	AA	Toxicity	102	Japanese	0.001	(Sugishita *et al.*, 2016)
Oxaliplatin And Fluorouracil	AG + GG	Increased Response	72	European	0.02	(Lamas *et al.*, 2011)
	AG + GG	Improved survival	72	European	0.015	(Lamas *et al.*, 2011)
	AA	Decreased survival	106	Mixed Population	0.019	(Stoehlmacher *et al.*, 2004)

The AA genotype has been found linked to increased risk of toxicity with platinum-based drugs in a variety of populations, but the same genotype has also been linked to positive effects of increased response of cyclophosphamide and epirubicin. Zhang *et al.* (2011) noted that along with the associated increased response, severity of toxic symptoms was decreased. The sum of these reports indicates that the A allele results in increased bioavailability of drugs, which may result in toxicity of platinum drugs, but also with better response to cyclophosphamide and epirubicin as the bioavailability of these drugs is increased.

The rs1695 AG and GG genotypes, although expected to lower drug bioavailability, have been linked to cisplatin induced ototoxicity (hearing loss). Drogemoller *et al.* (2017) also reported that adjusting drug dose had no effect on this association. The AG and GG genotypes have also been found to improve response to the combination of fluorouracil and oxaliplatin treatments, while the AA genotype is associated with treatment failure. This means that the detoxifying effects of GSTP1 may result in different outcomes based on which platinum drug is used in combination with fluorouracil.

As frequencies of both A and G alleles of rs1695 are common in Africans along with their respective genotype combinations, genotyping protocols could be put into place to guide treatments. The GG and AG genotype patients should be treated with the oxaliplatin/fluorouracil instead of cisplatin, whereas AA homozygotes should be treated with cyclophosphamide and epirubicin rather than platinum-based drugs, which are likely to induce toxic effects in these patients.

4.5 Fluoropyrimidine-related pharmacogenes

This section describes the relationship between variants in two genes which interact with a commonly used chemotherapy regimen – fluoropyrimidines.

4.5.1 Dihydropyrimidine Dehydrogenase - *DPYD*

DPYD is a critical pharmacogene as variants can produce extremely severe, lethally toxic effects (van Kuilenburg *et al.*, 2001), following treatment with fluorouracil based treatments. This gene contained the largest numbers of likely deleterious variants outside of those genes in the cytochrome P450 family, with 28 missense variants and one LoF variant. The rs72549310 stop gain variant was present only in the Esan population, with only one heterozygote present in the dataset. This variant has no current links to clinical findings, but has been shown induce a non-functional transcript in mammalian expression systems (Offer *et al.*, 2014). Though rare, rs72549310 is still relevant to pharmacogenomic research, as the most well characterised *DPYD* LoF variant, rs3918290 - a splice donor variant, has been linked to severe toxic effects during fluorouracil treatment in individuals who are heterozygous for the variant (Ceric *et al.*, 2010; Joerger *et al.*, 2015). The rs3918290 variant, however, is present in European and Asian populations, but not in any African populations assessed in this study and others (Elraiyah *et al.*, 2017).

Of the missense variants in the *DPYD* gene, 12 variants were found likely deleterious, all of which are rare in Africans, with most frequencies below 1%. As missense variants may not completely knock out the function of the gene, these may not be as critical as the rare LoF variant in heterozygous configurations. Of the 12 variants, five are the common variant in more than one African population: rs114096998, rs115232898, rs142619737, rs72975710 and rs2297595, which hold more relevance for broad PGx strategies compare to those which are extremely uncommon.

The rs114096998 variant is African specific, present in all African populations though with significantly different allele frequencies between populations. This variant is more common in West Africans at frequencies between ~2% and 6%, while rarer in East and South Africans (0-2%). The T allele of this variant has been noted to reduce activity (79% of wild type) in a cohort of East Africans from Kenya and Somalia (Elraiyah *et al.*, 2017), though the opposite effects (increased activity) were noted in a functional assay for allele T (Offer *et al.*, 2014). These findings match the predictions, with Polyphen-2 predicting a benign effect, and SIFT predicting a deleterious result with low confidence, while CADD and FATHMM both rank the variant as

likely functional. This variant requires further research to establish its impact on *DPYD* activity, particularly in clinical trials which would be best focused in West Africans.

The rs115232898 variant is more evenly spread across African populations with an average frequency of ~3% but is absent from Ethiopian populations. It is an African-specific variant, present elsewhere only in the Puerto Rican population from KGP Admixed-Americans (<1%) who may have some African ancestry. The rs115232898 variant CT genotype has been significantly associated with fluorouracil toxicity in an Afro-Caribbean woman who presented with life-threatening toxic effects (Zaanan *et al.*, 2014) and in another case report of an African American patient also presenting severe toxic effects (Saif *et al.*, 2014). A follow up functional study to that of the Zanaan *et al.*2014 found that enzyme activity was reduced by 29% in the affected patient (Offer and Diasio, 2014). The CT genotype has also been associated with lower DPYD activity in a cohort of healthy African Americans exposed to fluorouracil (n=94, p=0.00027) (Offer *et al.*, 2013). These individuals have this variant due to their African ancestry, and the effects confirm the high confidence predictions made by the toolsets used in this study. Strong likelihood of function and severe effects in individuals of African ancestry urge studies in continental Africans currently being treated with fluorouracil. Allele assays could therefore be used to place preventative guidelines in place to avoid toxicity.

The rs142619737 and rs72975710 variants have been found to produce no change in DPYD activity (Offer *et al.*, 2014), despite a high confidence prediction for rs142619737. These variants are both rare, with rs142619737 present only in the West African populations, and rs72975710 present in both the Nigerian populations, the Luhya and the Bagandan populations. Until further studies present any functional effect, these variants are most likely not useful for population pharmacogenomics guidance in Africans.

The rs2297595 variant is present in all African populations, but differs significantly in frequency between West, and East and South Africans, with the variant being rare in West Africans (~1%) and more common in East Africans (~8%) and South African Zulu (12%). It is significantly different between Africans and South and East Asians, and similar in frequency between Africans and Admixed Americans and Europeans. Predicted to be likely functionally deleterious by all toolsets, the rs2297595 CC and CT genotypes have been linked to fluorouracil toxicity in a cohort of European breast and gastroesophageal cancer patients (n=92, p=0.001) (Gross *et al.*, 2008) and a cohort of European colorectal cancer patients (n=568, p<0.05) (Deenen *et al.*, 2011). The CT genotype was also found associated with increased

severity of toxic effects in an Italian cohort of cancer patients (n=64, p0.022) (Falvella *et al.*, 2015). There have been conflicting reports that found no association with either genotype with toxicity in the following cohorts: a mixed population cohort (n=430, p=0.3957) (Loganayagam *et al.*, 2013); an Asian cohort of gastric cancer patients (n=362, p=4294) (Zhang *et al.*, 2012); and a case/control analysis of a European cohort of gastric cancer patients (n=410, 95 cases) (Toffoli *et al.*, 2015). These conflicting results encourage further research of this variant, particularly to assess linked variants that may be involved with the functional effect. This should be conducted primarily in East and South African populations, in which this variant is more common.

The seven other missense variants are all rare (<1%) and are each limited to only one population: COSM1688098, rs144395748, rs145548112, rs1801158, rs201268750, rs573299212 and rs60511679. The rs144395748, rs145548112, (both with high confidence consensus predictions) and rs60511679 variants were found to produce 92%, 87% and 86% activity of wild type DPYD activity (Offer *et al.*, 2014), and are thus not likely to affect treatment. The COSM1688098, rs201268750, rs573299212 are all high confidence predictions with all tools in consensus but have no reports of functional effect in clinical or functional settings, likely due to their extreme rarity, and each being presence in a single African population. The remaining variant, is the rs1801158 variant, which is present in Ethiopians at <1% frequency, but is more common in Europeans (3%), Admixed-Americans (~1%) and South Asians (~1%). The rs1801158 CT genotype has been linked to fluorouracil toxicity, although with conflicting reports. The variant was found linked to toxicity in a mixed population cohort 430 cancer patients treated with capecitabine (fluorouracil prodrug) (n=430, p=1E-16) (Loganayagam *et al.*, 2013), but multiple sources indicate no association with toxicity in a case/control study of European cancer patients (n=410, with 95 cases) (Toffoli *et al.*, 2015); a cohort of African American and European American cancer patients (n=173, 92 cases, p=0.15 (African Americans), p=0.89 (European Americans)) (Offer *et al.*, 2013); and in a large cohort of mixed population cancer patients (n=888, p=0.13) (Rosmarin *et al.*, 2014). The variant was not classified as functional by all toolsets, with Polyphen-2 predicting a benign variant, and Condel reporting Neutral due to the low Polyphen-2 score compared with the high score provided by SIFT. This variant may require more work to establish its role in other populations, but due to its rarity in Africa, and many clinical findings where it is not related to toxicity, it is not likely useful for population pharmacogenomics use in Africa.

Additional *DPYD* variants with clinical relevance but low scoring annotations are rs12022243 and rs17376848, which are intronic and synonymous variants respectively; and rs1801159, rs1801160, rs1801265, and rs61622928, which are all missense variants which failed score cut-offs for most toolsets. The rs12022243 variant is present in all African populations but is not significantly different in frequency between them. It is also common worldwide, present at an average frequency of 23%, and only significantly different between Africans (~17) and South Asians where it occurs in much higher frequencies (44%). The T allele has been associated (in a single study) with increased severity of toxicity from capecitabine treatment in European colorectal cancer patients (n=940, p=2.55E-5) (Rosmarin *et al.*, 2015). Further research of this variant is encouraged in African cancer patients who are presenting with toxic effects to better characterise the role of this intronic variant. The synonymous rs17376848 variant, has conflicting findings regarding its link to fluorouracil toxicity, with it being found linked to toxic effects in a cohort of European colorectal cancer patients (n=64, p=0.01) and no association with toxicity in a case/control Italian cohort (n=410, 95 cases) (Toffoli *et al.*, 2015). This variant is significantly rarer in Africans (~1%) compared to all other global populations, and with conflicting results in those populations where it is more common, it may not be useful for African pharmacogenomics guidance.

The rs1801159 and rs1801160 missense variants were scored as likely not functionally deleterious by most tools, and have clinical findings supporting these predictions, with no link to toxicity for either variant in multiple studies (Deenen *et al.*, 2011; Rosmarin *et al.*, 2014; Zhang *et al.*, 2012). The amino acid change is from valine to isoleucine, both highly similar amino acids. As these variants are common to all global populations, it is unlikely that they would show different effects in Africans to those in other populations. The rs61622928 variant has been noted in functional assays to decrease activity of DPYD (Ezzeldin *et al.*, 2005; Thomas *et al.*, 2016), but other studies present no effect on activity linked to this variant (Kuilenburg *et al.*, 2016). This variant is African specific, present in Africans at an average allele frequency of 9%, but present in no other global populations. Further research of this variant should aim to characterise the pharmacogenomic effect this variant which may have related to fluorouracil treatment.

The rs1801265 variant which induces an amino acid change from arginine to cysteine also was scored as low likelihood of deleterious outcome by most toolsets, as in the case of the *CYP2D6* rs16947 variant, an arginine to cysteine change was predicted to be likely benign, but had clinical reports for adverse drug reactions. This variant is present in all African populations,

with an average allele frequency of 43% which is significantly different to all other global populations, for which the frequency ranges from 9% in East Asians to 27% in South Asians. The rs1801265 variant has been found associated with toxicity and DPYD activity, but there are multiple conflicting findings. The AG and GG genotype have been associated with fluorouracil toxicity and shorter survival times in a cohort of children with lymphoblastic leukemia (n=147, $p < 0.05$) (Zhao *et al.* 2016), however, these findings have been directly contested, with the claim that the rationale used in Zhao *et al.*'s 2016 study may have led to incorrect results (Deenen *et al.*, 2017). Other sources found no rs1801265 genotype linked to toxicity in a cohort of European cancer patients (n=252, p=0.362) (Boisdron-Celle *et al.*, 2007); a large cohort of colorectal cancer patients (n=838, p=0.47) (Rosmarin *et al.*, 2014); and in a case/control analysis of gastrointestinal malignancies (n=131, 55 cases, $p > 0.05$) (Joerger *et al.*, 2015). These negative findings correlate with the predictions made by the predictive tools used in this study. The rs1801265 variant, although significantly more common in Africans, is not likely to induce deleterious effects on DPYD function, and thus unlikely to be involved in drug reactions involving fluorouracil.

4.5.2 Thymidylate Synthase - *TYMS*

TYMS had the least of all missense variants present before annotation, of which only one (rs559346114) was found to be likely deleterious. This variant is extremely rare in Africans, present only with a single allele in one Esan individual, and absent from all other global populations, which places significant doubt on its existence.

Only one variant present in Africans had clinical evidence - rs151264360 a 3' UTR region deletion. The homozygous deletion has been linked to decreased efficacy of capecitabine and fluorouracil based treatments in a large clinical trial consisting of mixed populations (n=738, p=0.012) (Rosmarin *et al.*, 2014). It was also found associated with shorter survival times following pemetrexed treatment in a small cohort of European mesothelioma patients (n=59, p=0.0185) (Powrozek *et al.*, 2014). The deletion variant is very common amongst Africans (~60%), compared to other global populations and is well described in terms of its effect, urging further study in African patients undergoing fluorouracil-based treatments.

4.5.3 Summary: Fluoropyrimidine-related pharmacogenes

Fluoropyrimidines are widely used for cancer treatment and can treat multiple types effectively, but are highly toxic, placing strong relevance on characterisation of variants in the gene encoding its primary metaboliser, *DYPD*. The majority of variants predicted to be deleterious in this gene are missense variants and they vary greatly in frequency among African

populations. The rs115232898 variant, which is African specific and produces severe toxicity, is absent in Ethiopians but present in all other African populations assessed. The rs2297595 variant which is known to produce toxicity effects (albeit with some conflicting reports), is more common in East and South Africans than in West African populations. Strategies aiming to guide treatment will need to take these population-based differences into account.

Strategies guiding fluoropyrimidine treatment would also benefit from characterising the *TYMS* rs151264360 variant, which can decrease drug efficacy, in addition to the potential toxic effects due to *DPYD* variants.

4.6 DNA Repair Proteins

This section describes two genes *XRCC1* and *XRCC5*, which repair DNA, counter-acting the action of specific cancer drug types.

4.6.1 X-Ray Repair Cross-Complementing 1 - *XRCC1*

The DNA repair gene, *XRCC1*, though having overall fewer variants than the other gene in its family assessed here, *XRCC5*, had a greater number of missense variants, as well as upstream gene variants. Of the 17 missense variants present across Africans in *XRCC1*, there were eight predicted to be likely deleterious, of which only one variant, rs1799782, has been annotated clinically. The rs1799782 variant was found to be predicted likely deleterious by almost all tools, only narrowly below the cut-off for FATHMM-MKL, but still in the deleterious score range for this tool. The rs1799782 AA and AG genotypes have been linked to better responses to platinum-based drugs, though with contradictory reports. Patients with the AA or AG genotypes were found to have a higher response rate to chemotherapy in an Asian cohort of non-small-cell lung cancer (n=200, $p<0.01$) (Liu *et al.*, 2013), but this finding was not present in a larger cohort, also of Asian non-small-cell lung cancer patients (n=375, $p=0.78$) (Zhang *et al.*, 2014). This variant is significantly rarer in Africans (7%) than in East Asians (28%), though no reports are available regarding its function in Africans. Though reports are conflicting regarding function, this variant should be investigated in African clinical trials as platinum-based therapies are widely used in the treatment of multiple cancer types.

Only one other missense variant which was predicted deleterious is common to all African populations, rs2307177. It is present at an average frequency of 5%, which is significantly different from all other global populations apart from Admixed-Americans. No clinical information is available for this variant, but further studies are recommended in Africans due to the absence of this variant in previously characterised populations.

One additional missense variant, rs25487, was included despite low scores from prediction tools. It has a Level 2 annotation for platinum-based drug types, the key findings for which are displayed in Table 4.6 below.

Table 4. 6 Key findings from literature for the high impact variant: rs25487

Drug	Alleles	Effect	Sample Size	Population	p-value	Reference
Platinum Based Drugs	TT+CT	Decreased survival	118	European	0.03	(Giovannetti *et al.*, 2011)
	TT+CT	Decreased survival	104	Siberian	p<0.05	(Khrunin *et al.*, 2012)
	TT	Decreased survival	118	European	0.001	(Kalikaki *et al.*, 2009)

The rs25487 TT and CT genotypes have been found to decrease overall survival following treatment with platinum-based drugs. This may mean that the rs2587 variant is not deleterious to the function of XRCC1, but may slightly improve its function instead, increasing repair of the DNA damage induced by platinum drugs, and thus the cancer will remain untreated. This variant is present in all African populations though with diverse frequencies. The T allele (minor allele) is more common in Bagandans (20%) and Ethiopians (18%) than in other African populations (mean frequency of 9%). These populations are more likely to have a poor response to platinum-based drugs. Globally, the T allele is significantly rarer in Africans than in European (37%), Admixed American (31%) East Asian (24%) and South Asian (34%) populations, meaning Africans in general are not as at risk for treatment failure with platinum-based drugs.

4.6.2 X-Ray Repair Cross Complementing 5 - *XRCC5*

XRCC5 had six missense variants present across the African populations, but none met scoring criteria to be predicted as likely deleterious. Two variants were annotated as likely deleterious by CADD and Fathmm, rs564358760 had high CADD and Fathmm scores, and rs548351594 had high CADD scores. Both are extremely rare, both present in only one population with frequencies below 0.5%. Neither of these variants has any clinical reports available.

One additional variant was found to be clinically significant: rs1051685, a 3'UTR region variant. The AG genotype was found to be linked to increased overall survival following

treatment with thalidomide in a small cohort of Spanish multiple myeloma patients (n=28, p=0.04) (Cibeira *et al.*, 2011). The G allele is more common in Africans than in other global populations, with an average frequency of 35%. This variant should be researched more in Africans to confirm this effect, as its higher frequencies indicate that Africans may respond better to thalidomide treatments than other global populations.

Chapter 5: Conclusions

This study characterised and annotated genetic variants in selected cancer treatment related PGx genes in WGS data from African populations. Annotation was performed using bioinformatics tools to predict the potential functional impact of variants, and predicted high impact variants were characterised by comparing their frequencies in African populations, and between Africans and other global populations. The following sections describe key findings and outcomes of the study.

5.1 PGx related findings to address cancer treatment in Africa

This study revealed numerous rare variants with strong predictions of functional impact using multiple *in silico* tools, yet in the absence of validation and clinical trials they remain of uncertain pharmacogenomic significance. In the first instance, the validity of the rare variants that were observed only once or twice should be investigated by performing Sanger sequencing or high coverage next generation sequencing to have the variant confirmed with confidence. Should they be confirmed, their clinical relevance would need to be assessed. This has not yet been done, primarily because they are extremely rare and have therefore not been considered significantly relevant to be included in research studies or clinical trials. This may be the case even for common variants which are African specific, since pharmacogenomics research is still developing on the continent.

Five novel African-specific variants of potential pharmacogenomic interest were identified (Table 3.3 and 3.4). This confirms the hypothesis that previously unknown variants of potential impact are present in African populations and can be detected by mining public databases with African data. These novel variants are however extremely rare, with a single allele presenting in one population and thus, as mentioned above, they need to be confirmed using deep sequencing. Two of the novel variants were observed more than once: *ABCB1* - 7:99270249 was observed four times in Ethiopians, and *CYP3A5* - chr7:87195405 was observed twice, in once in Yoruba and once in Ethiopians respectively. Although they may have an impact on drug treatment on an individual level, these variants are unlikely to be useful for public health PGx strategies due to their rarity.

In this study, many common variants with known functional significance regarding drug reactions were identified. Common variants are far more likely to be useful for public health based PGx. These common variants include LoF variants, which knock out gene function, and

missense variants, that could result in altered enzyme activity levels, as well as some non-coding/regulatory variants. Some showed patterns of allele frequency variation within Africa and often had marked allele frequency differences when compared to European and Asian populations. This has important implications for African patients using certain treatments, or treatment strategies designed for Europeans or Asians. Africans may have different drug responses to those expected based on their genomic variation. The frequencies are also different across African regions, so estimates based on a single population cannot be used as a proxy for the entire continent. Common variants of known effect are key to PGx strategies for public health. These are also best suited to guide clinical practice which ensures that patients receive the right drug type and dose, and which populations require genotyping prior to treatment.

For common variants of PGx interest, additional sequencing could be used to compliment the KGP and AGVP datasets by including additional populations from across the continent, along with functional validation in clinical groups currently treating African patients. More WGS data from Africa will increase the mining potential of PGx data.

As cancer rates are rising rapidly in Africa, significant differences in variant frequency in PGx genes at a population and individual level motivates that more research is required to ensure cancer drug safety. PGx guidance for cancer drugs will help to optimise drug dosage, prevent severe toxic reactions that may be lethal to the patient, and maximise treatment efficacy to increase post-treatment survival time. This must be prefaced with robust diagnostic testing systems available across the continent, which could be motivated for by the potential public health risk of ADRs.

Regarding the approach of the study, some key advantages of using a bioinformatics-based approach are that it provides an overview of genetics-based ADR risk relatively more readily than clinical trials, and it can be done with pre-existing sequencing data. This lowers the cost of foundation research and can help to guide clinical studies in future with a precision approach. However, some disadvantages of this approach need to be considered. The bioinformatics predictive tools are most useful and accurate for variants with large functional effects, such as LoF variants and missense variants that involve amino acids with large structural and charge differences. Unfortunately, this is not the case for the substitution of highly similar amino acids where tools often failed to accurately predict changes in function. This emphasises need for functional assays, as well as the need for complimentary databases of clinical PGx data such as PharmGKB in combination with bioinformatics methods to accurately assess the potential

functional impact of variants. PharmGKB is currently the most well curated and diverse database of PGx interaction data but is limited for African PGx use due to few studies having been carried out on continental Africans. The database is regularly updated with new findings, and would benefit greatly from more research conducted in diverse populations. Bias introduced by PharmGKB reported findings could be reduced by motivating for more studies which investigate novel variants in Africans.

5.2 Potential Limitations

Some limitations were identified throughout this study and include limitations related to the datasets, bioinformatics methods and PGx related implications.

Low coverage WGS data, although well suited for capturing common variation, is not ideal for the detection of rare, low frequency variants. For the rare variants identified as likely functionally deleterious in this study, it would be important to use deep sequencing to establish if they are real. This was beyond the scope of this study. Although the quality control steps undertaken in the KGP and AGVP studies are robust enough to remove poor quality calls, they have limitations when it comes to calling rare, low-frequency variants, since the QC approaches are probabilistic and require sufficient depth of coverage for accurate calling of specific variants.

In this study, the focus was on the assessment of coding variation because there are fewer tools that assess non-coding variation and they are currently limited in terms of the accuracy with which they can provide functional predictions. This study used only two tools which can predict the impact of non-coding variants, CADD and FATHMM-MKL. Due to the complex mechanisms in which these variants function, two scores alone are not sufficient to predict impact. As more tools are used, more confidence in the prediction is held via Bayesian inference. More tools and scores can also help to reduce the number of false positive hits. Non-coding variants were found in the PharmGKB database with significant clinical associations for drug reactions and it is interesting that although apparently validated by clinical findings, neither tool predicted these as likely functional. Another limitation of the tools: CADD, FATHMM-MKL, and Condel, is that they function via retrieval of scores from precomputed score datasets generated from the reference datasets available for human genomes. This means that any mismatches (such as a rare tri-allelic variant) can cause the tool to return no score for that variant. The reference sets are updated from time to time, and become increasingly complete and accurate, but current findings are limited to the quality of the reference dataset.

This would also have contributed to the tools not being able to score some of the variants which turned out to have clinical relevance.

Bioinformatics tools are a strong guide to identify variants that are likely functional, but the function of such variants must first be validated in terms of effect through clinical trials before they can be considered to be of PGx relevance. In addition, known PGx variants that are causal in other world populations may not have the exact same effect in Africans, as environmental and epistatic factors may cause African patients to respond differently.

5.3 Recommendations for future work

The following section contains suggestions for continuing this work, which will lead to more effective planning for and analysis of PGx in Africa.

Traditionally in pharmacogenomics research, genomic variation was related to pharmacokinetic and pharmacodynamic measures. As most studies linked patterns of variation (often haplotype groupings) within the respective gene to a measured outcome, set terms were used to denote the gene-drug relationship: i.e. *CYP2D6*29* (referring to a specific variant combination) is related to reduced enzymatic function, (slower metabolism of drug). Some haplotypes consisting of multiple variants are critically relevant for pharmacogenomics. As the toolsets used to predict likelihood of functionality are built for single variants only, haplotypes for series of variants are beyond the scope of this study. Obtaining frequencies for these requires substantial extra bioinformatics processes including phasing of the data, and are thus not reported for the populations assessed. Haplotypes are however highly relevant for PGx use, and will be considered in future investigations. Tools which could integrate or assess the effect of multiple variant haplotypes would be highly beneficial to predict their impact on gene function. Additional reports of frequencies of these haplotypes (not called in this study) could be coupled with potential effects to assess their public health impact. Additionally, whether a patient has one or two deleterious alleles in the same gene can affect the severity of the outcome. Assessing if the impact of variants in combination would be useful to investigate potential impact for individual genotypes.

For missense variants that are likely functional, *in silico* assessment can be taken further via protein structure modelling of these variants to establish if the variants will have a large impact on protein structure (Schmidt et al., 2014). These models can be used for drug interaction studies which can assess potential changes in binding energy or metabolic reactions, which can

guide clinical setting as well as provide the starting point for drug design. Tools such as MODELLER are readily available for this purpose (Webb and Sali, 2014).

Future work should include larger sample sizes, as well as including other population groups from the African continent. With more data available, and deeper sequencing, rare variants can be identified with certainty. This work and that of future mining studies should encourage more pharmacokinetic and pharmacodynamic trials to assess the impact of the variants identified. Once confirmed, these could be complemented with clinical trials focused on PGx based safety and efficacy in Africa.

5.4 Impact of this study

This study identified substantial variation in key PGx related genes for cancer treatment in Africa, and revealed variations in allele frequencies of known PGx related variants.

These findings should create awareness for pharmaceutical companies to perform more research to ensure safety and efficacy of their drugs for African patients. However, known PGx variants confirmed in other studies could be used as preliminary guides for treatment strategies in Africa, particularly for drugs that have severe and often fatal adverse reactions, such as fluorouracil.

As the genes assessed are each related to multiple drug types (which include those used for diseases other than cancer), this work also provides a basis for assessing PGx of drugs used to treat other major disease groups in Africa. Cardiovascular disease, human immunodeficiency virus infection and tuberculosis are all chronic conditions that require ongoing treatment, and drug choices could benefit from PGx guidance. There is a need for a centralised base of PGx data for African populations. Future research would benefit from a collection of African-focused work that would help to characterise the interactions of the diverse genomic variation of African populations with drugs that are commonly used.

6: References

Aboel Dahab, A., El-Hag, D., Moutamed, G.M., Aboel Dahab, S., Abuknesha, R., and Smith, N.W. (2016). Pharmacokinetic variations in cancer patients with liver dysfunction: applications and challenges of pharmacometabolomics. Cancer Chemother. Pharmacol.

Abraham, J.E., Guo, Q., Dorling, L., Tyrer, J., Ingle, S., Hardy, R., Vallier, A.-L., Hiller, L., Burns, R., Jones, L., et al. (2014). Replication of genetic polymorphisms reported to be associated with taxane-related sensory neuropathy in patients with early breast cancer treated with Paclitaxel. Clin. Cancer Res. *20*, 2466–2475.

Adeagbo, B.A., Bolaji, O.O., Olugbade, T.A., Durosinmi, M.A., Bolarinwa, R.A., and Masimirembwa, C. (2016). Influence of CYP3A5*3 and ABCB1 C3435T on clinical outcomes and trough plasma concentrations of imatinib in Nigerians with chronic myeloid leukaemia. J. Clin. Pharm. Ther. *41*, 546–551.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Arena, R., Guazzi, M., Lianov, L., Whitsel, L., Berra, K., Lavie, C.J., Kaminsky, L., Williams, M., Hivert, M.-F., Cherie Franklin, N., et al. (2015). Healthy lifestyle interventions to combat noncommunicable disease—a novel nonhierarchical connectivity model for key stakeholders. Eur. Heart J.

Bailey, S.D., Desai, K., Kron, K.J., Mazrooei, P., Sinnott-Armstrong, N.A., Treloar, A.E., Dowar, M., Thu, K.L., Cescon, D.W., Silvester, J., et al. (2016). Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. Nat. Genet. 48, 1260–1266.

Baselga, J., Campone, M., Piccart, M., Burris, H.A. 3rd, Rugo, H.S., Sahmoud, T., Noguchi, S., Gnant, M., Pritchard, K.I., Lebrun, F., et al. (2012). Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. N. Engl. J. Med. *366*, 520–529.

Benn-Torres, J., Bonilla, C., Robbins, C.M., Waterman, L., Moses, T.Y., Hernandez, W., Santos, E.R., Bennett, F., Aiken, W., Tullock, T., et al. (2008). Admixture and population stratification in African Caribbean populations. Ann. Hum. Genet. *72*, 90–98.

Beoris, M., Amos Wilson, J., Garces, J.A., and Lukowiak, A.A. (2016). CYP2D6 copy number distribution in the US population. Pharmacogenet. Genomics *26*, 96–99.

Birdwell, K.A., Grady, B., Choi, L., Xu, H., Bian, A., Denny, J.C., Jiang, M., Vranic, G., Basford, M., Cowan, J.D., et al. (2012). The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. Pharmacogenet. Genomics *22*, 32–42.

Bodor, M., Kelly, E.J., and Ho, R.J. (2005). Characterization of the human MDR1 gene. AAPS J. *7*, E1-5.

Boisdron-Celle, M., Remaud, G., Traore, S., Poirier, A.L., Gamelin, L., Morel, A., and Gamelin, E. (2007). 5-Fluorouracil-related severe toxicity: a comparison of different methods for the pretherapeutic detection of dihydropyrimidine dehydrogenase deficiency. Cancer Lett. *249*, 271–282.

Booton, R., Ward, T., Heighway, J., Ashcroft, L., Morris, J., and Thatcher, N. (2006). Glutathione-S-transferase P1 isoenzyme polymorphisms, platinum-based chemotherapy, and non-small cell lung cancer. J. Thorac. Oncol. *1*, 679–683.

Borobia, A.M., Romero, I., Jimenez, C., Gil, F., Ramirez, E., De Gracia, R., Escuin, F., Gonzalez, E.,

and Sansuan, A.J.C. (2009). Trough tacrolimus concentrations in the first week after kidney transplantation are related to acute rejection. Ther. Drug Monit. *31*, 436–442.

Boso, V., Herrero, M.J., Santaballa, A., Palomar, L., Megias, J.E., de la Cueva, H., Rojas, L., Marques, M.R., Poveda, J.L., Montalar, J., et al. (2014). SNPs and taxane toxicity in breast cancer patients. Pharmacogenomics *15*, 1845–1858.

Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.

Caronia, D., Patino-Garcia, A., Perez-Martinez, A., Pita, G., Moreno, L.T., Zalacain-Diez, M., Molina, B., Colmenero, I., Sierrasesumaga, L., Benitez, J., et al. (2011). Effect of ABCB1 and ABCC3 polymorphisms on osteosarcoma survival after chemotherapy: a pharmacogenetic study. PLoS One *6*, e26091.

Caudle, K.E., Dunnenberger, H.M., Freimuth, R.R., Peterson, J.F., Burlison, J.D., Whirl-Carrillo, M., Scott, S.A., Rehm, H.L., Williams, M.S., Klein, T.E., et al. (2017). Standardizing terms for clinical pharmacogenetic test results: Consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). Genet. Med. *19*, 215–223.

Ceric, T., Obralic, N., Kapur-Pojskic, L., Macic, D., Beslija, S., Pasic, A., and Ceric, S. (2010). Investigation of IVS14 + 1G > A polymorphism of DPYD gene in a group of Bosnian patients treated with 5-Fluorouracil and capecitabine. Bosn. J. Basic Med. Sci. *10*, 133–139.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chen, C.-H., Wang, S.-C., Tsou, H.-H., Ho, I.-K., Tian, J.-N., Yu, C.-J., Hsiao, C.-F., Chou, S.-Y., Lin, Y.-F., Fang, K.-C., et al. (2011). Genetic polymorphisms in CYP3A4 are associated with withdrawal symptoms and adverse reactions in methadone maintenance patients. Pharmacogenomics *12*, 1397–1406.

Chen, J., Wu, L., Wang, Y., Yin, J., Li, X., Wang, Z., Li, H., Zou, T., Qian, C., Li, C., et al. (2016). Effect of transporter and DNA repair gene polymorphisms to lung cancer chemotherapy toxicity. Tumour Biol. *37*, 2275–2284.

Chu, W., Fyles, A., Sellers, E.M., McCready, D.R., Murphy, J., Pal, T., and Narod, S.A. (2007). Association between CYP3A4 genotype and risk of endometrial cancer following tamoxifen use. Carcinogenesis *28*, 2139–2142.

Chu, Y.-H., Li, H., Tan, H.S., Koh, V., Lai, J., Phyo, W.M., Choudhury, Y., Kanesvaran, R., Chau, N.M., Toh, C.K., et al. (2015). Association of ABCB1 and FLT3 Polymorphisms with Toxicities and Survival in Asian Patients Receiving Sunitinib for Renal Cell Carcinoma. PLoS One *10*, e0134102.

Chung, H.H., Kim, M.-K., Kim, J.W., Park, N.-H., Song, Y.-S., Kang, S.-B., and Lee, H.-P. (2006). XRCC1 R399Q polymorphism is associated with response to platinum-based neoadjuvant chemotherapy in bulky cervical cancer. Gynecol. Oncol. *103*, 1031–1037.

Cibeira, M.T., de Larrea, C.F., Navarro, A., Diaz, T., Fuster, D., Tovar, N., Rosinol, L., Monzo, M., and Blade, J. (2011). Impact on response and survival of DNA repair single nucleotide polymorphisms in relapsed or refractory multiple myeloma patients treated with thalidomide. Leuk. Res. *35*, 1178–1183.

Ciccacci, C., Borgiani, P., Ceffa, S., Sirianni, E., Marazzi, M.C., Altan, A.M.D., Paturzo, G., Bramanti, P., Novelli, G., and Palombi, L. (2010). Nevirapine-induced hepatotoxicity and pharmacogenetics: a retrospective study in a population from Mozambique. Pharmacogenomics *11*, 23–31.

Colomer, R., Monzo, M., Tusquets, I., Rifa, J., Baena, J.M., Barnadas, A., Calvo, L., Carabantes, F.,

Crespo, C., Munoz, M., et al. (2008). A single-nucleotide polymorphism in the aromatase gene is associated with the efficacy of the aromatase inhibitor letrozole in advanced breast carcinoma. Clin. Cancer Res. *14*, 811–816.

Corrigan, A., Walker, J.L., Wickramasinghe, S., Hernandez, M.A., Newhouse, S.J., Folarin, A.A., Lewis, C.M., Sanderson, J.D., Spicer, J., and Marinaki, A.M. (2014). Pharmacogenetics of pemetrexed combination therapy in lung cancer: pathway analysis reveals novel toxicity associations. Pharmacogenomics J. *14*, 411–417.

Dahlman-Wright, K., Cavailles, V., Fuqua, S.A., Jordan, V.C., Katzenellenbogen, J.A., Korach, K.S., Maggi, A., Muramatsu, M., Parker, M.G., and Gustafsson, J.-A. (2006). International Union of Pharmacology. LXIV. Estrogen receptors. Pharmacol. Rev. *58*, 773–781.

Danenberg, P. V, Malli, H., and Swenson, S. (1999). Thymidylate synthase inhibitors. Semin. Oncol. *26*, 621–631.

de-Graft Aikins, A., Unwin, N., Agyemang, C., Allotey, P., Campbell, C., and Arhinful, D. (2010). Tackling Africa's chronic disease burden: from the local to the global. Global. Health *6*, 5.

Deenen, M.J., Tol, J., Burylo, A.M., Doodeman, V.D., de Boer, A., Vincent, A., Guchelaar, H.-J., Smits, P.H.M., Beijnen, J.H., Punt, C.J.A., et al. (2011). Relationship between single nucleotide polymorphisms and haplotypes in DPYD and toxicity and efficacy of capecitabine in advanced colorectal cancer. Clin. Cancer Res. *17*, 3455–3468.

Deenen, M.J., Henricks, L.M., Sonke, G.S., Schellens, J.H., and Meulendijks, D. (2017). Letter regarding Zhao et al. entitled " DPYD gene polymorphisms are associated with risk and chemotherapy prognosis in pediatric patients with acute lymphoblastic leukemia". Tumour Biol. *39*, 1010428317701629.

Depeille, P., Cuq, P., Passagne, I., Evrard, A., and Vian, L. (2005). Combined effects of GSTP1 and MRP1 in melanoma drug resistance. Br. J. Cancer *93*, 216–223.

Dickens, C., Duarte, R., Zietsman, A., Cubasch, H., Kellett, P., Schuz, J., Kielkowski, D., and McCormack, V. (2014). Racial comparison of receptor-defined breast cancer in Southern African women: subtype prevalence and age-incidence analysis of nationwide cancer registry data. Cancer Epidemiol. Biomarkers Prev. *23*, 2311–2321.

Dodgen, T.M., Hochfeld, W.E., Fickl, H., Asfaha, S.M., Durandt, C., Rheeder, P., Drögemöller, B.I., Wright, G.E.B., Warnich, L., Labuschagne, C.D.J., et al. (2013). Introduction of the AmpliChip CYP450 Test to a South African cohort: a platform comparative prospective cohort study. BMC Med. Genet. *14*, 20.

Drogemoller, B.I., Monzon, J.G., Bhavsar, A.P., Borrie, A.E., Brooks, B., Wright, G.E.B., Liu, G., Renouf, D.J., Kollmannsberger, C.K., Bedard, P.L., et al. (2017). Association Between SLC16A5 Genetic Variation and Cisplatin-Induced Ototoxic Effects in Adult Patients With Testicular Cancer. JAMA Oncol. *3*, 1558–1562.

Dunham, I., and Kundaje, A. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Egbelakin, A., Ferguson, M.J., MacGill, E.A., Lehmann, A.S., Topletz, A.R., Quinney, S.K., Li, L., McCammack, K.C., Hall, S.D., and Renbarger, J.L. (2011). Increased risk of vincristine neurotoxicity associated with low CYP3A5 expression genotype in children with acute lymphoblastic leukemia. Pediatr. Blood Cancer *56*, 361–367.

Elraiyah, T., Jerde, C.R., Shrestha, S., Wu, R., Nie, Q., Giama, N.H., Sarangi, V., Roberts, L.R., Offer, S.M., and Diasio, R.B. (2017). Novel Deleterious Dihydropyrimidine Dehydrogenase Variants May Contribute to 5-Fluorouracil Sensitivity in an East African Population. Clin. Pharmacol. Ther. *101*, 382–390.

Eng, A., McCormack, V., and dos-Santos-Silva, I. (2014). Receptor-defined subtypes of breast cancer in indigenous populations in Africa: a systematic review and meta-analysis. PLoS Med. *11*, e1001720.

Ezzeldin, H.H., Lee, A.M., Mattison, L.K., and Diasio, R.B. (2005). Methylation of the DPYD promoter: an alternative mechanism for dihydropyrimidine dehydrogenase deficiency in cancer patients. Clin. Cancer Res. *11*, 8699–8705.

Falvella, F.S., Cheli, S., Martinetti, A., Mazzali, C., Iacovelli, R., Maggi, C., Gariboldi, M., Pierotti, M.A., Di Bartolomeo, M., Sottotetti, E., et al. (2015). DPD and UGT1A1 deficiency in colorectal cancer patients receiving triplet chemotherapy with fluoropyrimidines, oxaliplatin and irinotecan. Br. J. Clin. Pharmacol. *80*, 581–588.

Findlay, J.M., Middleton, M.R., and Tomlinson, I. (2015). A systematic review and meta-analysis of somatic and germline DNA sequence biomarkers of esophageal cancer survival, therapy response and stage. Ann. Oncol. *26*, 624–644.

Fontein, D.B.Y., Houtsma, D., Nortier, J.W.R., Baak-Pablo, R.F., Kranenbarg, E.M.-K., van der Straaten, T.R.J.H.M., Putter, H., Seynaeve, C., Gelderblom, H., van de Velde, C.J.H., et al. (2014). Germline variants in the CYP19A1 gene are related to specific adverse events in aromatase inhibitor users: a substudy of Dutch patients in the TEAM trial. Breast Cancer Res. Treat. *144*, 599–606.

Di Francia, R., De Lucia, L., Di Paolo, M., Di Martino, S., Del Pup, L., De Monaco, A., Lleshi, A., and Berretta, M. (2015). Rational selection of predictive pharmacogenomics test for the Fluoropyrimidine/Oxaliplatin based therapy. Eur. Rev. Med. Pharmacol. Sci. *19*, 4443–4454.

Garcia-Donas, J., Esteban, E., Leandro-Garcia, L.J., Castellano, D.E., del Alba, A.G., Climent, M.A., Arranz, J.A., Gallardo, E., Puente, J., Bellmunt, 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. Lancet. Oncol. *12*, 1143–1150.

Gervasini, G., Carrillo, J.A., Garcia, M., San Jose, C., Cabanillas, A., and Benitez, J. (2006). Adenosine triphosphate-binding cassette B1 (ABCB1) (multidrug resistance 1) G2677T/A gene polymorphism is associated with high risk of lung cancer. Cancer *107*, 2850–2857.

Gervasini, G., Jara, C., Olier, C., Romero, N., Martinez, R., and Carrillo, J.A. (2017a). Polymorphisms in ABCB1 and CYP19A1 genes affect anastrozole plasma concentrations and clinical outcomes in postmenopausal breast cancer patients. Br. J. Clin. Pharmacol. *83*, 562–571.

Gervasini, G., Garcia-Pino, G., Vergara, E., Mota-Zamorano, S., Garcia-Cerrada, M., and Luna, E. (2017b). CYP3A genotypes of donors but not those of the patients increase the risk of acute rejection in renal transplant recipients on calcineurin inhibitors: a pilot study. Eur. J. Clin. Pharmacol.

Giovannetti, E., Pacetti, P., Reni, M., Leon, L.G., Mambrini, A., Vasile, E., Ghidini, M., Funel, N., Lucchesi, M., Cereda, S., et al. (2011). Association between DNA-repair polymorphisms and survival in pancreatic cancer patients treated with combination chemotherapy. Pharmacogenomics *12*, 1641–1652.

Goetz, M.P., Rae, J.M., Suman, V.J., Safgren, S.L., Ames, M.M., Visscher, D.W., Reynolds, C., Couch, F.J., Lingle, W.L., Flockhart, D.A., et al. (2005). Pharmacogenetics of tamoxifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. J. Clin. Oncol. *23*, 9312–9318.

Goff, L.W., Thakkar, N., Du, L., Chan, E., Tan, B.R., Cardin, D.B., McLeod, H.L., Berlin, J.D., Zehnbauer, B., Fournier, C., et al. (2014). Thymidylate synthase genotype-directed chemotherapy for patients with gastric and gastroesophageal junction cancers. PLoS One *9*, e107424.

Gonen, N., and Assaraf, Y.G. (2012). Antifolates in cancer therapy: structure, activity and mechanisms of drug resistance. Drug Resist. Updat. *15*, 183–210.

González-Pérez, A., and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. *88*, 440–449.

Goricar, K., Kovac, V., Jazbec, J., Zakotnik, B., Lamovec, J., and Dolzan, V. (2015). Genetic variability of DNA repair mechanisms and glutathione-S-transferase genes influences treatment outcome in osteosarcoma. Cancer Epidemiol. *39*, 182–188.

Green, H., Falk, I.J., Lotfi, K., Paul, E., Hermansson, M., Rosenquist, R., Paul, C., and Nahi, H. (2012). Association of ABCB1 polymorphisms with survival and in vitro cytotoxicty in de novo acute myeloid leukemia with normal karyotype. Pharmacogenomics J. *12*, 111–118.

Gregers, J., Green, H., Christensen, I.J., Dalhoff, K., Schroeder, H., Carlsen, N., Rosthoej, S., Lausen, B., Schmiegelow, K., and Peterson, C. (2015). Polymorphisms in the ABCB1 gene and effect on outcome and toxicity in childhood acute lymphoblastic leukemia. Pharmacogenomics J. *15*, 372–379.

Gross, E., Ullrich, T., Seck, K., Mueller, V., de Wit, M., von Schilling, C., Meindl, A., Schmitt, M., and Kiechle, M. (2003). Detailed analysis of five mutations in dihydropyrimidine dehydrogenase detected in cancer patients with 5-fluorouracil-related side effects. Hum. Mutat. *22*, 498.

Gross, E., Busse, B., Riemenschneider, M., Neubauer, S., Seck, K., Klein, H.-G., Kiechle, M., Lordick, F., and Meindl, A. (2008). Strong association of a common dihydropyrimidine dehydrogenase gene polymorphism with fluoropyrimidine-related toxicity in cancer patients. PLoS One *3*, e4003.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.

Haas, D.W., Ribaudo, H.J., Kim, R.B., Tierney, C., Wilkinson, G.R., Gulick, R.M., Clifford, D.B., Hulgan, T., Marzolini, C., and Acosta, E.P. (2004). Pharmacogenetics of efavirenz and central nervous system side effects: an Adult AIDS Clinical Trials Group study. AIDS *18*, 2391–2400.

Haas, D.W., Bartlett, J.A., Andersen, J.W., Sanne, I., Wilkinson, G.R., Hinkle, J., Rousseau, F., Ingram, C.D., Shaw, A., Lederman, M.M., et al. (2006). Pharmacogenetics of nevirapine-associated hepatotoxicity: an Adult AIDS Clinical Trials Group collaboration. Clin. Infect. Dis. *43*, 783–786.

Han, K.-M., Chang, H.S., Choi, I.-K., Ham, B.-J., and Lee, M.-S. (2013). CYP2D6 P34S Polymorphism and Outcomes of Escitalopram Treatment in Koreans with Major Depression. Psychiatry Investig. *10*, 286–293.

Harivenkatesh, N., Kumar, L., Bakhshi, S., Sharma, A., Kabra, M., Velpandian, T., Gogia, A., Shastri, S.S., and Gupta, Y.K. (2017). Do polymorphisms in MDR1 and CYP3A5 genes influence the risk of cytogenetic relapse in patients with chronic myeloid leukemia on imatinib therapy? Leuk. Lymphoma *58*, 1–9.

Ben Hassine, I., Gharbi, H., Soltani, I., Ben Hadj Othman, H., Farrah, A., Amouri, H., Teber, M., Ghedira, H., Ben Youssef, Y., Safra, I., et al. (2017). Molecular study of ABCB1 gene and its correlation with imatinib response in chronic myeloid leukemia. Cancer Chemother. Pharmacol. *80*, 829–839.

He, H.-R., Liu, P., He, G.-H., Dong, W.-H., Wang, M.-Y., Dong, Y.-L., and Lu, J. (2014). Association between reduced folate carrier G80A polymorphism and methotrexate toxicity in childhood acute lymphoblastic leukemia: a meta-analysis. Leuk. Lymphoma *55*, 2793–2800.

Henry, N.L., Skaar, T.C., Dantzer, J., Li, L., Kidwell, K., Gersch, C., Nguyen, A.T., Rae, J.M., Desta, Z., Oesterreich, S., et al. (2013). Genetic associations with toxicity-related discontinuation of aromatase inhibitor therapy for breast cancer. Breast Cancer Res. Treat. *138*, 807–816.

Hertz, D.L., Motsinger-Reif, A.A., Drobish, A., Winham, S.J., McLeod, H.L., Carey, L.A., and Dees,

E.C. (2012). CYP2C8*3 predicts benefit/risk profile in breast cancer patients receiving neoadjuvant paclitaxel. Breast Cancer Res. Treat. *134*, 401–410.

Hesselink, D.A., van Schaik, R.H.N., van der Heiden, I.P., van der Werf, M., Gregoor, P.J.H.S., Lindemans, J., Weimar, W., and van Gelder, T. (2003). Genetic polymorphisms of the CYP3A4, CYP3A5, and MDR-1 genes and pharmacokinetics of the calcineurin inhibitors cyclosporine and tacrolimus. Clin. Pharmacol. Ther. *74*, 245–254.

Hicks, J.K., Bishop, J.R., Sangkuhl, K., Müller, D.J., Ji, Y., Leckband, S.G., Leeder, J.S., Graham, R.L., Chiulli, D.L., LLerena, A., et al. (2015). Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and CYP2C19 Genotypes and Dosing of Selective Serotonin Reuptake Inhibitors. Clin. Pharmacol. Ther. *98*, 127–134.

Hodges, L., Markova, S.M., Chinn, L.W., Gow, J.M., Kroetz, D.L., Klein, T.E., and Altman, R.B. (2012). A very important pharmacogene summary: ABCB1 (MDR1, P-glycoportien). Pharmacogenet Genomics *21*, 152–161.

Hu, L., Lv, Q.-L., Guo, Y., Cheng, L., Wu, N.-Y., Qin, C.-Z., and Zhou, H.-H. (2016). Genetic variation of CYP3A5 influences paclitaxel/carboplatin-induced toxicity in Chinese epithelial ovarian cancer patients. J. Clin. Pharmacol. *56*, 349–354.

Huang, Y. (2007). Pharmacogenetics/genomics of membrane transporters in cancer chemotherapy. Cancer Metastasis Rev. *26*, 183–201.

Huang, M.-Y., Huang, M.-L., Chen, M.-J., Lu, C.-Y., Chen, C.-F., Tsai, P.-C., Chuang, S.-C., Hou, M.-F., Lin, S.-R., and Wang, J.-Y. (2011). Multiple genetic polymorphisms in the prediction of clinical outcome of metastatic colorectal cancer patients treated with first-line FOLFOX-4 chemotherapy. Pharmacogenet. Genomics *21*, 18–25.

Imanishi, H., Okamura, N., Yagi, M., Noro, Y., Moriya, Y., Nakamura, T., Hayakawa, A., Takeshima, Y., Sakaeda, T., Matsuo, M., et al. (2007). Genetic polymorphisms associated with adverse events and elimination of methotrexate in childhood acute lymphoblastic leukemia and malignant lymphoma. J. Hum. Genet. *52*, 166–171.

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

Jacobson, P.A., Oetting, W.S., Brearley, A.M., Leduc, R., Guan, W., Schladt, D., Matas, A.J., Lamba, V., Julian, B.A., Mannon, R.B., et al. (2011). Novel polymorphisms associated with tacrolimus trough concentrations: results from a multicenter kidney transplant consortium. Transplantation *91*, 300–308.

Jansson, I., Stoilov, I., Sarfarazi, M., and Schenkman, J.B. (2001). Effect of two mutations of human CYP1B1, G61E and R469W, on stability and endogenous steroid substrate metabolism. Pharmacogenetics *11*, 793–801.

Ji, Y., Schaid, D.J., Desta, Z., Kubo, M., Batzler, A.J., Snyder, K., Mushiroda, T., Kamatani, N., Ogburn, E., Hall-Flavin, D., et al. (2014). Citalopram and escitalopram plasma drug and metabolite concentrations: genome-wide associations. Br. J. Clin. Pharmacol. *78*, 373–383.

Joerger, M., Huitema, A.D.R., Boot, H., Cats, A., Doodeman, V.D., Smits, P.H.M., Vainchtein, L., Rosing, H., Meijerman, I., Zueger, M., et al. (2015). Germline TYMS genotype is highly predictive in patients with metastatic gastrointestinal malignancies receiving capecitabine-based chemotherapy. Cancer Chemother. Pharmacol. *75*, 763–772.

Jones, M.E.E., Boon, W.C., McInnes, K., Maffei, L., Carani, C., and Simpson, E.R. (2007). Recognizing rare disorders: aromatase deficiency. Nat. Clin. Pract. Endocrinol. Metab. *3*, 414–421.

de Jonge, H., Metalidis, C., Naesens, M., Lambrechts, D., and Kuypers, D.R.J. (2011). The P450 oxidoreductase *28 SNP is associated with low initial tacrolimus exposure and increased dose requirements in CYP3A5-expressing renal recipients. Pharmacogenomics *12*, 1281–1291.

Jordheim, L.P., Ribrag, V., Ghesquieres, H., Pallardy, S., Delarue, R., Tilly, H., Haioun, C., Jardin, F., Demangel, D., Salles, G.A., et al. (2015). Single nucleotide polymorphisms in ABCB1 and CBR1 can predict toxicity to R-CHOP type regimens in patients with diffuse non-Hodgkin lymphoma. Haematologica *100*, e204-6.

Kalman, L. V, Agúndez, J.A.G., Appell, M.L., Black, J.L., Bell, G.C., Boukouvala, S., Bruckner, C., Bruford, E., Bruckner, C., Caudle, K., et al. (2016). Pharmacogenetic Allele Nomenclature: International Workgroup Recommendations for Test Result Reporting. Clin. Pharmacol. Ther. 99, 172–185.

Kagimoto, M., Heim, M., Kagimoto, K., Zeugin, T., and Meyer, U.A. (1990). Multiple mutations of the human cytochrome P450IID6 gene (CYP2D6) in poor metabolizers of debrisoquine. Study of the functional significance of individual mutations by expression of chimeric genes. J. Biol. Chem. *265*, 17209–17214.

Kalikaki, A., Kanaki, M., Vassalou, H., Souglakos, J., Voutsina, A., Georgoulias, V., and Mavroudis, D. (2009). DNA repair gene polymorphisms predict favorable clinical outcome in advanced non-small-cell lung cancer. Clin. Lung Cancer *10*, 118–123.

Kehdy, F.S.G., Gouveia, M.H., Machado, M., Magalhaes, W.C.S., Horimoto, A.R., Horta, B.L., Moreira, R.G., Leal, T.P., Scliar, M.O., Soares-Souza, G.B., et al. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc. Natl. Acad. Sci. U. S. A. *112*, 8696–8701.

Khrunin, A., Ivanova, F., Moisseev, A., Khokhrin, D., Sleptsova, Y., Gorbunova, V., and Limborska, S. (2012). Pharmacogenomics of cisplatin-based chemotherapy in ovarian cancer patients of different ethnic origins. Pharmacogenomics *13*, 171–178.

Kim, D.H., Park, J.Y., Sohn, S.K., Lee, N.Y., Baek, J.H., Jeon, S.B., Kim, J.G., Suh, J.S., Do, Y.R., and Lee, K.B. (2006). Multidrug resistance-1 gene polymorphisms associated with treatment outcomes in de novo acute myeloid leukemia. Int. J. Cancer *118*, 2195–2201.

Kim, H.R., Park, H.S., Kwon, W.S., Lee, J.H., Tanigawara, Y., Lim, S.M., Kim, H.S., Shin, S.J., Ahn, J.B., and Rha, S.Y. (2013). Pharmacogenetic determinants associated with sunitinib-induced toxicity and ethnic difference in Korean metastatic renal cell carcinoma patients. Cancer Chemother. Pharmacol. *72*, 825–835.

Kim, H.S., Kim, M.-K., Chung, H.H., Kim, J.W., Park, N.H., Song, Y.S., and Kang, S.B. (2009). Genetic polymorphisms affecting clinical outcomes in epithelial ovarian cancer patients treated with taxanes and platinum compounds: a Korean population-based study. Gynecol. Oncol. *113*, 264–269.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

de Klerk, O.L., Nolte, I.M., Bet, P.M., Bosker, F.J., Snieder, H., den Boer, J.A., Bruggeman, R., Hoogendijk, W.J., and Penninx, B.W. (2013). ABCB1 gene variants influence tolerance to selective serotonin reuptake inhibitors in a large sample of Dutch cases with major depressive disorder. Pharmacogenomics J. *13*, 349–353.

Kobayashi, H., Sato, K., Niioka, T., Miura, H., Ito, H., and Miura, M. (2015). Relationship Among Gefitinib Exposure, Polymorphisms of Its Metabolizing Enzymes and Transporters, and Side Effects in Japanese Patients With Non-Small-Cell Lung Cancer. Clin. Lung Cancer *16*, 274–281.

Kuilenburg, A.B.P. van, Meijer, J., Tanck, M.W.T., Dobritzsch, D., Zoetekouw, L., Dekkers, L.-L., Roelofsen, J., Meinsma, R., Wymenga, M., Kulik, W., et al. (2016). Phenotypic and clinical implications of variants in the dihydropyrimidine dehydrogenase gene. Biochim. Biophys. Acta *1862*, 754–762.

van Kuilenburg, A.B.P., Muller, E.W., Haasjes, J., Meinsma, R., Zoetekouw, L., Waterham, H.R., Baas, F., Richel, D.J., and van Gennip, A.H. (2001). Lethal Outcome of a Patient with a Complete Dihydropyrimidine Dehydrogenase (DPD) Deficiency after Administration of 5-Fluorouracil. Clin. Cancer Res. *7*, 1149 LP-1153.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

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

Kus, T., Aktas, G., Kalender, M.E., Demiryurek, A.T., Ulasli, M., Oztuzcu, S., Sevinc, A., Kul, S., and Camci, C. (2016). Polymorphism of CYP3A4 and ABCB1 genes increase the risk of neuropathy in breast cancer patients treated with paclitaxel and docetaxel. Onco. Targets. Ther. *9*, 5073–5080.

Lal, S., Wong, Z.W., Sandanaraj, E., Xiang, X., Ang, P.C.S., Lee, E.J.D., and Chowbay, B. (2008). Influence of ABCB1 and ABCG2 polymorphisms on doxorubicin disposition in Asian breast cancer patients. Cancer Sci. *99*, 816–823.

Lamas, M.J., Duran, G., Balboa, E., Bernardez, B., Touris, M., Vidal, Y., Gallardo, E., Lopez, R., Carracedo, A., and Barros, F. (2011). Use of a comprehensive panel of biomarkers to predict response to a fluorouracil-oxaliplatin regimen in patients with metastatic colorectal cancer. Pharmacogenomics *12*, 433–442.

Langaee, T., Hamadeh, I., Chapman, A.B., Gums, J.G., and Johnson, J.A. (2015). A novel simple method for determining CYP2D6 gene copy number and identifying allele(s) with duplication/multiplication. PLoS One *10*, e0113808.

Ledesma, M.C., and Agundez, J.A.G. (2005). Identification of subtypes of CYP2D gene rearrangements among carriers of CYP2D6 gene deletion and duplication. Clin. Chem. *51*, 939–943.

Lee, K.-H., Chang, H.J., Han, S.-W., Oh, D.-Y., Im, S.-A., Bang, Y.-J., Kim, S.Y., Lee, K.-W., Kim, J.H., Hong, Y.S., et al. (2013). Pharmacogenetic analysis of adjuvant FOLFOX for Korean patients with colon cancer. Cancer Chemother. Pharmacol. *71*, 843–851.

Lee, S.-Y., Im, S.-A., Park, Y.H., Woo, S.Y., Kim, S., Choi, M.K., Chang, W., Ahn, J.S., and Im, Y.-H. (2014). Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. Eur. J. Cancer *50*, 698–705.

Lee, W., Lockhart, A.C., Kim, R.B., and Rothenberg, M.L. (2005). Cancer pharmacogenomics: powerful tools in cancer chemotherapy and drug development. Oncologist *10*, 104–111.

Lek, M., Karczewski, K.J., Minikel, E. V, Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285.

Leskela, S., Jara, C., Leandro-Garcia, L.J., Martinez, A., Garcia-Donas, J., Hernando, S., Hurtado, A., Vicario, J.C.C., Montero-Conde, C., Landa, I., et al. (2011). Polymorphisms in cytochromes P450 2C8 and 3A5 are associated with paclitaxel neurotoxicity. Pharmacogenomics J. *11*, 121–129.

Li, S., Shen, D., Shao, J., Crowder, R., Liu, W., Prat, A., He, X., Liu, S., Hoog, J., Lu, C., et al. (2013). Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. Cell Rep. *4*, 1116–1130.

Lima, A., Azevedo, R., Sousa, H., Seabra, V., and Medeiros, R. (2013). Current approaches for TYMS polymorphisms and their importance in molecular epidemiology and pharmacogenetics. Pharmacogenomics *14*, 1337–1351.

Liu, L., Wu, J., Zhong, R., Wu, C., Zou, L., Yang, B., Chen, W., Zhu, B., Duan, S., Yu, D., et al. (2013). Multi-loci analysis reveals the importance of genetic variations in sensitivity of platinum-based chemotherapy in non-small-cell lung cancer. Mol. Carcinog. *52*, 923–931.

Liu, S.-G., Gao, C., Zhang, R.-D., Zhao, X.-X., Cui, L., Li, W.-J., Chen, Z.-P., Yue, Z.-X., Zhang, Y.-Y., Wu, M.-Y., et al. (2017). Polymorphisms in methotrexate transporters and their relationship to plasma methotrexate levels, toxicity of high-dose methotrexate, and outcome of pediatric acute lymphoblastic leukemia. Oncotarget *8*, 37761–37772.

Loganayagam, A., Arenas Hernandez, M., Corrigan, A., Fairbanks, L., Lewis, C.M., Harper, P., Maisey, N., Ross, P., Sanderson, J.D., and Marinaki, A.M. (2013). Pharmacogenetic variants in the DPYD, TYMS, CDA and MTHFR genes are clinically significant predictors of fluoropyrimidine toxicity. Br. J. Cancer *108*, 2505–2515.

Lopes, M.C., Joyce, C., Ritchie, G.R.S., John, S.L., Cunningham, F., Asimit, J., and Zeggini, E. (2012). A Combined Functional Annotation Score for Non-Synonymous Variants. Hum. Hered. *73*, 47–51.

Lunenburg, C.A.T.C., Henricks, L.M., Guchelaar, H.-J., Swen, J.J., Deenen, M.J., Schellens, J.H.M., and Gelderblom, H. (2016). Prospective DPYD genotyping to reduce the risk of fluoropyrimidine-induced severe toxicity: Ready for prime time. Eur. J. Cancer *54*, 40–48.

Lyon, E., Gastier Foster, J., Palomaki, G.E., Pratt, V.M., Reynolds, K., Sábato, M.F., Scott, S. a, and Vitazka, P. (2012). Laboratory testing of CYP2D6 alleles in relation to tamoxifen therapy. Genet. Med. *14*, 990–1000.

Ma, Y., Xin, S., Huang, M., Yang, Y., Zhu, C., Zhao, H., Zhang, Y., Chen, L., Zhao, Y., Li, J., et al. (2017). Determinants of Gefitinib toxicity in advanced non-small cell lung cancer (NSCLC): a pharmacogenomic study of metabolic enzymes and transporters. Pharmacogenomics J. *17*, 325–330.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

Mahon, F.-X., Belloc, F., Lagarde, V., Chollet, C., Moreau-Gaudry, F., Reiffers, J., Goldman, J.M., and Melo, J. V (2003). MDR1 gene overexpression confers resistance to imatinib mesylate in leukemia cell line models. Blood *101*, 2368 LP-2373.

Maldonado, A.Q., Asempa, T., Hudson, S., and Rebellato, L.M. (2017). Prevalence of CYP3A5 Genomic Variances and Their Impact on Tacrolimus Dosing Requirements among Kidney Transplant Recipients in Eastern North Carolina. Pharmacotherapy *37*, 1081–1088.

Mayosi, B.M., Flisher, A.J., Lalloo, U.G., Sitas, F., Tollman, S.M., and Bradshaw, D. (2009). The burden of non-communicable diseases in South Africa. Lancet *374*, 934–947.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

McLeod, H.L., Sargent, D.J., Marsh, S., Green, E.M., King, C.R., Fuchs, C.S., Ramanathan, R.K., Williamson, S.K., Findlay, B.P., Thibodeau, S.N., et al. (2010). Pharmacogenetic predictors of adverse events and response to chemotherapy in metastatic colorectal cancer: results from North American Gastrointestinal Intergroup Trial N9741. J. Clin. Oncol. *28*, 3227–3233.

Megias-Vericat, J.E., Rojas, L., Herrero, M.J., Boso, V., Montesinos, P., Moscardo, F., Poveda, J.L., Sanz, M.A., and Alino, S.F. (2015). Influence of ABCB1 polymorphisms upon the effectiveness of standard treatment for acute myeloid leukemia: a systematic review and meta-analysis of observational studies. Pharmacogenomics J. *15*, 109–118.

Meibohm, B., and Derendorf, H. (1997). Basic concepts of pharmacokinetic/pharmacodynamic (PK/PD) modelling. Int J Clin Pharmacol Ther. *35*, 401-413

Meulendijks, D., Jacobs, B.A.W., Aliev, A., Pluim, D., van Werkhoven, E., Deenen, M.J., Beijnen, J.H., Cats, A., and Schellens, J.H.M. (2016). Increased risk of severe fluoropyrimidine-associated toxicity in patients carrying a G to C substitution in the first 28-bp tandem repeat of the thymidylate synthase 2R allele. Int. J. Cancer *138*, 245–253.

Morel, A., Boisdron-Celle, M., Fey, L., Soulie, P., Craipeau, M.C., Traore, S., and Gamelin, E. (2006). Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide polymorphisms on 5-fluorouracil tolerance. Mol. Cancer Ther. *5*, 2895–2904.

Le Morvan, V., Litiere, S., Laroche-Clary, A., Ait-Ouferoukh, S., Bellott, R., Messina, C., Cameron, D., Bonnefoi, H., and Robert, J. (2015). Identification of SNPs associated with response of breast cancer patients to neoadjuvant chemotherapy in the EORTC-10994 randomized phase III trial. Pharmacogenomics J. *15*, 63–68.

Murray, T., Beaty, T.H., Mathias, R.A., Rafaels, N., Grant, A.V., Faruque, M.U., Watson, H.R., Ruczinski, I., Dunston, G.M., and Barnes, K.C. (2010). African and Non-African Admixture Components in African Americans and An African Caribbean Population. Genet. Epidemiol. *34*, 10.1002/gepi.20512.

Mutagonda, R.F., Kamuhabwa, A.A.R., Minzi, O.M.S., Massawe, S.N., Asghar, M., Homann, M. V, Farnert, A., and Aklillu, E. (2017). Effect of pharmacogenetics on plasma lumefantrine pharmacokinetics and malaria treatment outcome in pregnant women. Malar. J. *16*, 267.

Nakajima, Y., Yoshitani, T., Fukushima-Uesaka, H., Saito, Y., Kaniwa, N., Kurose, K., Ozawa, S., Aoyagi, N., Kamatani, N., Yamamoto, N., et al. (2006). Impact of the haplotype CYP3A4*16B harboring the Thr185Ser substitution on paclitaxel metabolism in Japanese patients with cancer. Clin. Pharmacol. Ther. *80*, 179–191.

Narumiya, K., Metzger, R., Bollschweiler, E., Alakus, H., Brabender, J., Drebber, U., Holscher, A.H., and Warnecke-Eberz, U. (2011). Impact of ABCB1 C3435T polymorphism on lymph node regression in multimodality treatment of locally advanced esophageal cancer. Pharmacogenomics *12*, 205–214.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Niioka, T., Satoh, S., Kagaya, H., Numakura, K., Inoue, T., Saito, M., Narita, S., Tsuchiya, N., Habuchi, T., and Miura, M. (2012). Comparison of pharmacokinetics and pharmacogenetics of once- and twice-daily tacrolimus in the early stage after renal transplantation. Transplantation *94*, 1013–1019.

Ntukidem, N.I., Nguyen, A.T., Stearns, V., Rehman, M., Schott, A., Skaar, T., Jin, Y., Blanche, P., Li, L., Lemler, S., et al. (2008). Estrogen receptor genotypes, menopausal status, and the lipid effects of tamoxifen. Clin. Pharmacol. Ther. *83*, 702–710.

O'Brien, F.E., Dinan, T.G., Griffin, B.T., and Cryan, J.F. (2012). Interactions between antidepressants and P-glycoprotein at the blood-brain barrier: clinical significance of in vitro and in vivo findings. Br. J. Pharmacol. *165*, 289–312.

Oesterreich, S., Henry, N.L., Kidwell, K.M., Van Poznak, C.H., Skaar, T.C., Dantzer, J., Li, L., Hangartner, T.N., Peacock, M., Nguyen, A.T., et al. (2015). Associations between genetic variants and the effect of letrozole and exemestane on bone mass and bone turnover. Breast Cancer Res. Treat. *154*, 263–273.

Offer, S.M., and Diasio, R.B. (2014). Response to "A case of 5-FU-related severe toxicity associated with the P.Y186C DPYD variant". Clin. Pharmacol. Ther. *95*, 137.

Offer, S.M., Lee, A.M., Mattison, L.K., Fossum, C., Wegner, N.J., and Diasio, R.B. (2013). A DPYD variant (Y186C) in individuals of african ancestry is associated with reduced DPD enzyme activity. Clin. Pharmacol. Ther. *94*, 158–166.

Offer, S.M., Fossum, C.C., Wegner, N.J., Stuflesser, A.J., Butterfield, G.L., and Diasio, R.B. (2014). Comparative functional analysis of DPYD variants of potential clinical relevance to dihydropyrimidine dehydrogenase activity. Cancer Res. *74*, 2545–2554.

Ohtsuki, S., Schaefer, O., Kawakami, H., Inoue, T., Liehner, S., Saito, A., Ishiguro, N., Kishimoto, W., Ludwig-Schwellinger, E., Ebner, T., et al. (2012). Simultaneous absolute protein quantification of transporters, cytochromes P450, and UDP-glucuronosyltransferases as a novel approach for the characterization of individual human liver: comparison with mRNA levels and activities. Drug Metab. Dispos. *40*, 83–92.

Oliveira, A.L., Rodrigues, F.F.O., Santos, R.E., Aoki, T., Rocha, M.N., Longui, C.A., and Melo, M.B. (2010). GSTT1, GSTM1, and GSTP1 polymorphisms and chemotherapy response in locally advanced breast cancer. Genet. Mol. Res. *9*, 1045–1053.

Otyepka, M., Berka, K., and Anzenbacher, P. (2012). Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450? Curr. Drug Metab. 13, 130–142.

Owen, S.A., Hider, S.L., Martin, P., Bruce, I.N., Barton, A., and Thomson, W. (2013). Genetic polymorphisms in key methotrexate pathway genes are associated with response to treatment in rheumatoid arthritis patients. Pharmacogenomics J. *13*, 227–234.

Papamichael, D. (1999). The use of thymidylate synthase inhibitors in the treatment of advanced colorectal cancer: current status. Oncologist *4*, 478–487.

Pascual, T., Apellaniz-Ruiz, M., Pernaut, C., Cueto-Felgueroso, C., Villalba, P., Alvarez, C., Manso, L., Inglada-Perez, L., Robledo, M., Rodriguez-Antona, C., et al. (2017). Polymorphisms associated with everolimus pharmacokinetics, toxicity and survival in metastatic breast cancer. PLoS One *12*, e0180192.

Passey, C., Birnbaum, A.K., Brundage, R.C., Oetting, W.S., Israni, A.K., and Jacobson, P.A. (2011). Dosing equation for tacrolimus using genetic variants and clinical factors. Br. J. Clin. Pharmacol. *72*, 948–957.

Patel, J.N. (2015). Cancer pharmacogenomics: implications on ethnic diversity and drug response. Pharmacogenet. Genomics *25*, 1.

Pellicer, M., Garcia-Gonzalez, X., Garcia, M.I., Robles, L., Gravalos, C., Garcia-Alfonso, P., Pachon, V., Longo, F., Martinez, V., Blanco, C., et al. (2017). Identification of new SNPs associated with severe toxicity to capecitabine. Pharmacol. Res. *120*, 133–137.

Pettan-Brewer, C., Morton, J., Cullen, S., Enns, L., Kehrli, K.R., Sidorova, J., Goh, J., Coil, R., and Ladiges, W.C. (2012). Tumor growth is suppressed in mice expressing a truncated XRCC1 protein. Am. J. Cancer Res. *2*, 168–177.

Pirmohamed, M. (2001) Pharmacogenetics and Pharmacogenomics. Br J Clin Pharmacol. *54*, 345-347

Powrozek, T., Kowalski, D.M., Krawczyk, P., Ramlau, R., Kucharczyk, T., Kalinka-Warzocha, E., Knetki-Wroblewska, M., Winiarczyk, K., Dyszkiewicz, W., Krzakowski, M., et al. (2014). Correlation between TS, MTHFR, and ERCC1 gene polymorphisms and the efficacy of platinum in combination with pemetrexed first-line chemotherapy in mesothelioma patients. Clin. Lung Cancer *15*, 455–465.

Preissner, S.C., Hoffmann, M.F., Preissner, R., Dunkel, M., Gewiess, A., and Preissner, S. (2013). Polymorphic Cytochrome P450 Enzymes (CYPs) and Their Role in Personalized Therapy. PLoS One *8*, e82562.

Quteineh, L., Verstuyft, C., Furlan, V., Durrbach, A., Letierce, A., Ferlicot, S., Taburet, A.-M., Charpentier, B., and Becquemont, L. (2008). Influence of CYP3A5 genetic polymorphism on tacrolimus daily dose requirements and acute rejection in renal graft recipients. Basic Clin. Pharmacol. Toxicol. *103*, 546–552.

R Core Team (2015). R: A Language and Environment for Statistical Computing.

Rajman, I., Knapp, L., Morgan, T., and Masimirembwa, C. (2017). African Genetic Diversity: Implications for Cytochrome P450-mediated Drug Metabolism and Drug Development. EBioMedicine *17*, 67–74.

Ranganathan, P., Culverhouse, R., Marsh, S., Mody, A., Scott-Horton, T.J., Brasington, R., Joseph, A., Reddy, V., Eisen, S., and McLeod, H.L. (2008). Methotrexate (MTX) pathway gene polymorphisms and their effects on MTX toxicity in Caucasian and African American patients with rheumatoid arthritis. J. Rheumatol. *35*, 572–579.

Rebsamen, M., Desmeules, J., Daali, Y., Chiappe, A., Diemand, A., Rey, C., Chabert, J., Dayer, P., Hochstrasser, D., and Rossier, M. (2009). The AmpliChip CYP450 test: Cytochrome P450 2D6 genotype assessment and phenotype prediction.

Rednam, S., Scheurer, M.E., Adesina, A., Lau, C.C., and Okcu, M.F. (2013). Glutathione S-transferase P1 single nucleotide polymorphism predicts permanent ototoxicity in children with medulloblastoma. Pediatr. Blood Cancer *60*, 593–598.

Richardson, T.G., Campbell, C., Timpson, N.J., and Gaunt, T.R. (2016). Incorporating Non-Coding Annotations into Rare Variant Analysis. PLoS One *11*, e0154181.

Ritchie, M.D., Haas, D.W., Motsinger, A.A., Donahue, J.P., Erdem, H., Raffanti, S., Rebeiro, P., George, A.L., Kim, R.B., Haines, J.L., et al. (2006). Drug transporter and metabolizing enzyme gene variants and nonnucleoside reverse-transcriptase inhibitor hepatotoxicity. Clin. Infect. Dis. *43*, 779–782.

Rosmarin, D., Palles, C., Church, D., Domingo, E., Jones, A., Johnstone, E., Wang, H., Love, S., Julier, P., Scudder, C., et al. (2014). Genetic markers of toxicity from capecitabine and other fluorouracil-based regimens: investigation in the QUASAR2 study, systematic review, and meta-analysis. J. Clin. Oncol. *32*, 1031–1039.

Rosmarin, D., Palles, C., Pagnamenta, A., Kaur, K., Pita, G., Martin, M., Domingo, E., Jones, A., Howarth, K., Freeman-Mills, L., 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 *64*, 111–120.

Rudd, G.N., Hartley, J.A., and Souhami, R.L. (1995). Persistence of cisplatin-induced DNA interstrand crosslinking in peripheral blood mononuclear cells from elderly and young individuals. Cancer Chemother. Pharmacol. *35*, 323–326.

Rudin, C.M., Durinck, S., Stawiski, E.W., Poirier, J.T., Modrusan, Z., Shames, D.S., Bergbower, E.A., Guan, Y., Shin, J., Guillory, J., et al. (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat. Genet. *44*, 1111–1116.

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*, 89–92.

Saif, M.W., Lee, A.M., Offer, S.M., McConnell, K., Relias, V., and Diasio, R.B. (2014). A DPYD variant (y186c) specific to individuals of african descent in a patient with life-threatening 5-fu toxic effects: Potential for an individualized medicine approach. Mayo Clin. Proc. *89*, 131–136.

Saladores, P., Murdter, T., Eccles, D., Chowbay, B., Zgheib, N.K., Winter, S., Ganchev, B., Eccles, B., Gerty, S., Tfayli, A., et al. (2015). Tamoxifen metabolism predicts drug concentrations and outcome in premenopausal patients with early breast cancer. Pharmacogenomics J. *15*, 84–94.

Salazar, J., Altes, A., del Rio, E., Estella, J., Rives, S., Tasso, M., Navajas, A., Molina, J., Villa, M., Vivanco, J.L., et al. (2012). Methotrexate consolidation treatment according to pharmacogenetics of MTHFR ameliorates event-free survival in childhood acute lymphoblastic leukaemia.

Pharmacogenomics J. *12*, 379–385.

Samavat, H., and Kurzer, M.S. (2015). Estrogen Metabolism and Breast Cancer. Cancer Lett. *356*, 231–243.

Santa-Maria, C.A., Blackford, A., Nguyen, A.T., Skaar, T.C., Philips, S., Oesterreich, S., Rae, J.M., Desta, Z., Robarge, J., Henry, N.L., et al. (2016). Association of Variants in Candidate Genes with Lipid Profiles in Women with Early Breast Cancer on Adjuvant Aromatase Inhibitor Therapy. Clin. Cancer Res. *22*, 1395–1402.

Santoro, A., Felipe, C.R., Tedesco-Silva, H., Medina-Pestana, J.O., Struchiner, C.J., Ojopi, E.B., and Suarez-Kurtz, G. (2011). Pharmacogenetics of calcineurin inhibitors in Brazilian renal transplant patients. Pharmacogenomics *12*, 1293–1303.

Santoro, A.B., Struchiner, C.J., Felipe, C.R., Tedesco-Silva, H., Medina-Pestana, J.O., and Suarez-Kurtz, G. (2013). CYP3A5 genotype, but not CYP3A4*1b, CYP3A4*22, or hematocrit, predicts tacrolimus dose requirements in Brazilian renal transplant patients. Clin. Pharmacol. Ther. *94*, 201–202.

Sawers, L., Ferguson, M.J., Ihrig, B.R., Young, H.C., Chakravarty, P., Wolf, C.R., and Smith, G. (2014). Glutathione S-transferase P1 (GSTP1) directly influences platinum drug chemosensitivity in ovarian tumour cell lines. Br. J. Cancer *111*, 1150–1158.

Schmidt, T., Bergner, A., and Schwede, T. (2014). Modelling three-dimensional protein structures for applications in drug design. Drug Discov. Today *19*, 890–897.

Schroth, W., Antoniadou, L., Fritz, P., Schwab, M., Muerdter, T., Zanger, U.M., Simon, W., Eichelbaum, M., and Brauch, H. (2007). Breast cancer treatment outcome with adjuvant tamoxifen relative to patient CYP2D6 and CYP2C19 genotypes. J. Clin. Oncol. *25*, 5187–5193.

Schroth, W., Goetz, M.P., Hamann, U., Fasching, P.A., Schmidt, M., Winter, S., Fritz, P., Simon, W., Suman, V.J., Ames, M.M., et al. (2009). Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. JAMA *302*, 1429–1436.

Shang, B., Jia, Y., Chen, G., and Wang, Z. (2017). Ku80 correlates with neoadjuvant chemotherapy resistance in human lung adenocarcinoma, but reduces cisplatin/pemetrexed-induced apoptosis in A549 cells. Respir. Res. *18*, 56.

Shao, X., Guo, Y., Xu, X., Zheng, Y., Wang, J., Chen, Z., Huang, J., Huang, P., Cai, J., and Wang, X. (2015a). The CYP19 RS4646 polymorphism IS related to the prognosis of stage I-II and operable stage III breast cancer. PLoS One *10*, e0121535.

Shao, X., Cai, J., Zheng, Y., Wang, J., Feng, J., Huang, Y., Shi, L., Chen, Z., Guo, Y., and Wang, X. (2015b). S4646 polymorphism in CYP19A1 gene is associated with the efficacy of hormone therapy in early breast cancer. Int. J. Clin. Exp. Pathol. *8*, 5309–5317.

Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65.

Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536–1543.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. *15*, 121–132.

Singh, J.A. (2013). Is HLA B genotyping the future of gout pharmacogenomics? Nat. Rev. Rheumatol. *9*, 200–202.

Sissung, T.M., Danesi, R., Price, D.K., Steinberg, S.M., de Wit, R., Zahid, M., Gaikwad, N.,

Cavalieri, E., Dahut, W.L., Sackett, D.L., et al. (2008). Association of the CYP1B1*3 allele with survival in patients with prostate cancer receiving docetaxel. Mol. Cancer Ther. *7*, 19–26.

Standish, K.A., Carland, T.M., Lockwood, G.K., Pfeiffer, W., Tatineni, M., Huang, C.C., Lamberth, S., Cherkas, Y., Brodmerkel, C., Jaeger, E., et al. (2015). Group-based variant calling leveraging next-generation supercomputing for large-scale whole-genome sequencing studies. BMC Bioinformatics *16*, 304.

Stenberg, G., Abdalla, A.M., and Mannervik, B. (2000). Tyrosine 50 at the subunit interface of dimeric human glutathione transferase P1-1 is a structural key residue for modulating protein stability and catalytic function. Biochem. Biophys. Res. Commun. 271, 59–63.

Stingl, J.C., Parmar, S., Huber-Wechselberger, A., Kainz, A., Renner, W., Seeringer, A., Brockmoller, J., Langsenlehner, U., Krippl, P., and Haschke-Becher, E. (2010). Impact of CYP2D6*4 genotype on progression free survival in tamoxifen breast cancer treatment. Curr. Med. Res. Opin. *26*, 2535–2542.

Stoehlmacher, J., Park, D.J., Zhang, W., Yang, D., Groshen, S., Zahedy, S., and Lenz, H.-J. (2004). A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. Br. J. Cancer *91*, 344–354.

Su, H.I., Sammel, M.D., Velders, L., Horn, M., Stankiewicz, C., Matro, J., Gracia, C.R., Green, J., and DeMichele, A. (2010). Association of cyclophosphamide drug-metabolizing enzyme polymorphisms and chemotherapy-related ovarian failure in breast cancer survivors. Fertil. Steril. *94*, 645–654.

Sugishita, M., Imai, T., Kikumori, T., Mitsuma, A., Shimokata, T., Shibata, T., Morita, S., Inada-Inoue, M., Sawaki, M., Hasegawa, Y., et al. (2016). Pharmacogenetic association between GSTP1 genetic polymorphism and febrile neutropenia in Japanese patients with early breast cancer. Breast Cancer *23*, 195–201.

Suthandiram, S., Gan, G.-G., Zain, S.M., Bee, P.-C., Lian, L.-H., Chang, K.-M., Ong, T.-C., and Mohamed, Z. (2014). Effect of polymorphisms within methotrexate pathway genes on methotrexate toxicity and plasma levels in adults with hematological malignancies. Pharmacogenomics *15*, 1479–1494.

Tang, N.L.S., Liao, C. Di, Wang, X., Mo, F.K.F., Chan, V.T.C., Ng, R., Pang, E., Suen, J.J.S., Woo, J., and Yeo, W. (2013). Role of pharmacogenetics on adjuvant chemotherapy-induced neutropenia in Chinese breast cancer patients. J. Cancer Res. Clin. Oncol. *139*, 419–427.

Tavira, B., Coto, E., Diaz-Corte, C., Alvarez, V., Lopez-Larrea, C., and Ortega, F. (2013). A search for new CYP3A4 variants as determinants of tacrolimus dose requirements in renal-transplanted patients. Pharmacogenet. Genomics *23*, 445–448.

The 1000 Genomes Project Consortium, 2015, Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Thomas, F., Hennebelle, I., Delmas, C., Lochon, I., Dhelens, C., Garnier Tixidre, C., Bonadona, A., Penel, N., Goncalves, A., Delord, J.P., et al. (2016). Genotyping of a family with a novel deleterious DPYD mutation supports the pretherapeutic screening of DPD deficiency with dihydrouracil/uracil ratio. Clin. Pharmacol. Ther. *99*, 235–242.

Toffoli, G., Giodini, L., Buonadonna, A., Berretta, M., De Paoli, A., Scalone, S., Miolo, G., Mini, E., Nobili, S., Lonardi, S., et al. (2015). Clinical validity of a DPYD-based pharmacogenetic test to predict severe toxicity to fluoropyrimidines. Int. J. Cancer *137*, 2971–2980.

Townsend, D.M., and Tew, K.D. (2003). The role of glutathione-S-transferase in anti-cancer drug resistance. Oncogene *22*, 7369–7375.

Tulsyan, S., Chaturvedi, P., Agarwal, G., Lal, P., Agrawal, S., Mittal, R.D., and Mittal, B. (2013).

Pharmacogenetic influence of GST polymorphisms on anthracycline-based chemotherapy responses and toxicity in breast cancer patients: a multi-analytical approach. Mol. Diagn. Ther. *17*, 371–379.

Valente, C., Alvarez, L., Marks, S.J., Lopez-Parra, A.M., Parson, W., Oosthuizen, O., Oosthuizen, E., Amorim, A., Capelli, C., Arroyo-Pardo, E., et al. (2015). Exploring the relationship between lifestyles, diets and genetic adaptations in humans. BMC Genet. *16*, 55.

de Vries Schultink, A.H.M., Zwart, W., Linn, S.C., Beijnen, J.H., and Huitema, A.D.R. (2015). Effects of Pharmacogenetics on the Pharmacokinetics and Pharmacodynamics of Tamoxifen. Clin. Pharmacokinet. *54*, 797–810.

Walker, J.R., Corpina, R.A., and Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature *412*, 607–614.

Wang, D., Poi, M.J., Sun, X., Gaedigk, A., Leeder, J.S., and Sadee, W. (2014). Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. Hum. Mol. Genet. *23*, 268–278.

Wang, J., Lu, K., Song, Y., Xie, L., Zhao, S., Wang, Y., Sun, W., Liu, L., Zhao, H., Tang, D., et al. (2013). Indications of clinical and genetic predictors for aromatase inhibitors related musculoskeletal adverse events in Chinese Han women with breast cancer. PLoS One *8*, e68798.

Wang, L., Ellsworth, K.A., Moon, I., Pelleymounter, L.L., Eckloff, B.W., Martin, Y.N., Fridley, B.L., Jenkins, G.D., Batzler, A., Suman, V.J., et al. (2010). Functional genetic polymorphisms in the aromatase gene CYP19 vary the response of breast cancer patients to neoadjuvant therapy with aromatase inhibitors. Cancer Res. *70*, 319–328.

Warnich, L., Drögemöller, B.I., Pepper, M.S., Dandara, C., Wright, G.E.B., Town, C., and Africa, S. (2011). Pharmacogenomic Research in South Africa : Lessons Learned and Future Opportunities in the Rainbow Nation. 191–207.

Webb, B., and Sali, A. (2014). Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinforma. *47*, 5.6.1-32.

Wennerholm, A., Johansson, I., Hidestrand, M., Bertilsson, L., Gustafsson, L.L., and Ingelman-Sundberg, M. (2001). Characterization of the CYP2D6*29 allele commonly present in a black Tanzanian population causing reduced catalytic activity. Pharmacogenetics *11*, 417–427.

Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B., and Klein, T.E. (2012). Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther. *92*, 414–417.

WHO (2014). Global status report on noncommunicable diseases 2014 (Geneva, Switzerland: WHO Press).

Wilson, P.M., Danenberg, P. V, Johnston, P.G., Lenz, H.-J., and Ladner, R.D. (2014). Standing the test of time: targeting thymidylate biosynthesis in cancer therapy. Nat. Rev. Clin. Oncol. *11*, 282–298.

Winner, J., Allen, J.D., Anthony Altar, C., and Spahic-Mihajlovic, A. (2013). Psychiatric pharmacogenomics predicts health resource utilization of outpatients with anxiety and depression. Transl. Psychiatry *3*, e242.

Wright, G.E.B., Niehaus, D.J.H., Drogemuller, B.I., Koen, L., Gaedigk, A., and Warnich, L. (2010). Elucidation of CYP2D6 genetic diversity in a unique African population: Implications for the future application of pharmacogenetics in the xhosa population. Ann. Hum. Genet. *74*, 340–350.

Yao, S., Sucheston, L.E., Zhao, H., Barlow, W.E., Zirpoli, G., Liu, S., Moore, H.C.F., Thomas Budd, G., Hershman, D.L., Davis, W., et al. (2014). Germline genetic variants in ABCB1, ABCC1 and ALDH1A1, and risk of hematological and gastrointestinal toxicities in a SWOG Phase III trial S0221 for breast cancer. Pharmacogenomics J. *14*, 241–247.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44*, D710-6.

Yu, C.-C., Huang, S.-P., Lee, Y.-C., Huang, C.-Y., Liu, C.-C., Hour, T.-C., Huang, C.-N., You, B.-J., Chang, T.-Y., Huang, C.-H., et al. (2013). Molecular markers in sex hormone pathway genes associated with the efficacy of androgen-deprivation therapy for prostate cancer. PLoS One *8*, e54627.

Yuan, H., Yu, M., Yang, Y., Wu, K., Lin, X., and Li, J. (2010). Association of CYP2D6 single-nucleotide polymorphism with response to ophthalmic timolol in primary open-angle Glaucoma--a pilot study. J. Ocul. Pharmacol. Ther. *26*, 497–501.

Yuan, Z., Li, J., Hu, R., Jiao, Y., Han, Y., and Weng, Q. (2015). Predictive assessment in pharmacogenetics of XRCC1 gene on clinical outcomes of advanced lung cancer patients treated with platinum-based chemotherapy. Sci. Rep. *5*, 16482.

Zaanan, A., Dumont, L.-M., Loriot, M.-A., Taieb, J., and Narjoz, C. (2014). A case of 5-FU-related severe toxicity associated with the p.Y186C DPYD variant. Clin. Pharmacol. Ther. *95*, 136.

Zanger, U.M., Turpeinen, M., Klein, K., and Schwab, M. (2008). Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. Anal. Bioanal. Chem. *392*, 1093–1108.

Zgheib, N.K., Akra-Ismail, M., Aridi, C., Mahfouz, R., Abboud, M.R., Solh, H., and Muwakkit, S.A. (2014). Genetic polymorphisms in candidate genes predict increased toxicity with methotrexate therapy in Lebanese children with acute lymphoblastic leukemia. Pharmacogenet. Genomics *24*, 387–396.

Zhang, B.-L., Sun, T., Zhang, B.-N., Zheng, S., Lu, N., Xu, B.-H., Wang, X., Chen, G.-J., Yu, D.-K., and Lin, D.-X. (2011). Polymorphisms of GSTP1 is associated with differences of chemotherapy response and toxicity in breast cancer. Chin. Med. J. (Engl). *124*, 199–204.

Zhang, L., Ma, W., Li, Y., Wu, J., and Shi, G.Y. (2014). Pharmacogenetics of DNA repair gene polymorphisms in non-small-cell lung carcinoma patients on platinum-based chemotherapy. Genet. Mol. Res. *13*, 228–236.

Zhang, X., Bai, Z., Chen, B., Feng, J., Yan, F., Jiang, Z., Zhong, Y., Wu, J., Chen, L., Lu, Z., et al. (2012). Polymorphisms of dihydropyrimidine dehydrogenase gene and clinical outcomes of gastric cancer patients treated with fluorouracil-based adjuvant chemotherapy in Chinese population. Chin. Med. J. (Engl). *125*, 741–746.

Zhao, R., and Goldman, I.D. (2013). Folate and thiamine transporters mediated by facilitative carriers (SLC19A1-3 and SLC46A1) and folate receptors. Mol. Aspects Med. *34*, 373–385.

Zhao, X.-Q., Cao, W.-J., Yang, H.-P., Yang, X.-W., Tang, P., Sun, L., and Gao, X. (2016). DPYD gene polymorphisms are associated with risk and chemotherapy prognosis in pediatric patients with acute lymphoblastic leukemia. Tumour Biol.

Zheng, S., Tasnif, Y., Hebert, M.F., Davis, C.L., Shitara, Y., Calamia, J.C., Lin, Y.S., Shen, D.D., and Thummel, K.E. (2012). Measurement and compartmental modeling of the effect of CYP3A5 gene variation on systemic and intrarenal tacrolimus disposition. Clin. Pharmacol. Ther. *92*, 737–745.

7: Appendices

7.1 Documentation

Appendix 1: Ethics Documentation

Human Research Ethics Committee (Medical)

Research Office Secretariat: Senate House Room SH10005, 10th floor. Tel +27 (0)11-717-1252
Medical School Secretariat: P V Tobias Health Sciences Building, 2rd floor
Tel +27 (0)11-717-2700 / 1234 / 1252 / 2656
Private Bag 3, Wits 2050, www.wits.ac.za. Fax +27 (0)11-717-1265
South African National Health Research Ethics Council registration: REC-250208-04
United States Office of Human Research Protections registration: FWA00000715 IRB00001223

Ref: W-CJ-160815-2 15/08/2016

TO WHOM IT MAY CONCERN:

Waiver: This certifies that the following research does not require clearance from the Human Research Ethics Committee (Medical).

Investigator: **Jorge da Rocha (Student No 670830)**

Project title: **Addressing cancer treatment in an African setting: a bioinformatics analysis of pharmacogenomically relevant variants.**

Reason: This is a secondary analysis of data in a computational database the databases are the 1000 Genome Project (KGP) and the African Genome Variation Project (AGVP). The KGP data are in the public domain. Permission to use AGVP data has been granted by the database gatekeeper. There are no human participants.

UNIVERSITY OF THE WITWATERSRAND
PROF PE CLEATON-JONES
HREC (MEDICAL)
2016 -08- 15
JOHANNESBURG

Professor Peter Cleaton-Jones

Chair: Human Research Ethics Committee (Medical)

Copy – HREC (Medical) Secretariat: Zanele Ndlovu, Rhulani Mkansi.

Appendix 2: Plagiarism Declaration and Report

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 Jorge da Rocha (Student number: 670830) am a student registered for the degree of Master of Science in Medicine in the academic year 2018.

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: 27-03-2018

1

670830:Jorge_da_Rocha_Msc_670830_final_draft_text_only....

ORIGINALITY REPORT

4%	2%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Witwatersrand Student Paper	<1%
2	El-Hattab, Ayman W., Ranad Shaheen, Jozef Hertecant, Hassan I. Galadari, Badi S. Albaqawi, Amira Nabil, and Fowzan S. Alkuraya. "On the phenotypic spectrum of serine biosynthesis defects", Journal of Inherited Metabolic Disease, 2016. Publication	<1%
3	Li Ou, Michael J. Przybilla, Chester B. Whitley. "Phenotype prediction for mucopolysaccharidosis type I by in silico analysis", Orphanet Journal of Rare Diseases, 2017 Publication	<1%
4	Ryota Teramoto, Noboru Fujino, Tetsuo Konno, Akihiro Nomura et al. "Late Gadolinium Enhancement for Prediction of Mutation-Positive Hypertrophic Cardiomyopathy on the Basis of Panel-Wide Sequencing", Circulation	<1%

7.2 Methodology additions

Appendix Table 1: Prediction scores interpretations for analysis tools

Tool	Score	Prediction	Interpretation
Variant Effect Predictor (VEP)	None	Annotates Variant Biotype	Distinguishes variants into classes
Polyphen-2	0.899 < x ≤ 1	Probably Damaging	Higher Confidence
	0.499 < x ≤ 0.899	Possibly Damaging	Lower Confidence
	0.000 ≤ x ≤0.499	Benign	
SIFT	0.00 ≤ x ≤ 0.09	Deleterious	High Confidence
	0.09 < x < 1.00	Tolerated	
CAROL	0.981 ≤ x	Deleterious	High Confidence
	0.981 ≥ x	Neutral	
Condel	0.455 ≤ x	Deleterious	High Confidence
	0.454 ≥ x	Neutral	
FATHMM-MKL	0.9 ≤ x < 1	High Confidence - Deleterious	Most Likely Functional
	0.7 ≤ x < 0.9	Lower Confidence - Deleterious	Likely Functional
	0.5 ≤ x < 0.7	Lowest Confidence - Deleterious	Least Likely Functional
	0.30 ≤ x < 0.5	Lowest Confidence - Benign	Least Likely Non. Functional
	0.10 ≤ x < 0.30	Lower Confidence - Benign	Likely Non. Functional
	0.00 ≤ x < 0.10	Highest Confidence - Benign	Most Likely Non. Functional
CADD	30 < x < 40	Top 0.1% Most Deleterious	Greatest Functional Impact
	20 < x < 30	Top 1% Most Deleterious	Greater Functional Impact
	10 < x < 20	Top 10% Most Deleterious	Functional Impact
	X < 10	Irrelevant Score	Unpredictive range 25% - 75%
LOFTEE	None	HC	High Confidence
		LC	Low Confidence

Appendix Table 2: List of genes and pharmacogenomically relevant variants arranged by cancer type used for motivation of selection

Gene Name	Cancer Types	Variant Reported	Type	Effect	Population Reported	Major drugs affected*	Citation
ABCB1	Oesophageal	rs1045642	Synonymous	Ile1145Ile	European	oxaliplatin fluorouracil leucovorin cisplatin	(Narumiya *et al.*, 2011)
	Breast	rs1128503 rs2032582 rs3213619 rs1045642	Synonymous Missense 5' UTR Synonymous	Gly412Gly Ser893Ala Ile1145Ile	Asian Asian Asian European/Asian	paclitaxel gemcitabine fluorouracil cytarabine idarubicin taxanes doxorubicin tamoxifen docetaxel anthracyclines cyclophosphamide epirubicin docetaxel gemcitabine cyclophosphamide epirubicin	(Lal *et al.*, 2008) (Narumiya *et al.*, 2011)
	Multiple Myeloma	rs1045642 rs2032582	Synonymous Missense	Ile1145Ile Ser893Ala	European/Asian European/Asian	anthracyclines dexamethasone cytarabine doxorubicin idarubicin epirubicin vincristine	(Megias-Vericat *et al.*, 2015)

DPYD	Breast	rs1801159 rs2297595 rs3918290 rs45589337	Missense Missense Splice Donor Missense	Ile543Val Met166Val Lys259Glu	European European Asian	fluorouracil capecitabine cyclophosphamide methotrexate	(Toffoli *et al.*, 2015) (Zhao *et al.*, 2016) (Gross *et al.*, 2003)
TYMS	Oesophageal Breast	rs34743033 rs45445694 rs34743033 rs183205964	28bp repeat 28bp repeat 28bp repeat Substitution	5' UTR 5' UTR 5' UTR 5' UTR	European/Asian European/Asian European/Asian Mixed	leucovorin oxaliplatin fluorouracil gemcitabine carboplatin cyclophosphamide fluorouracil	(Goff *et al.*, 2014) (Meulendijks *et al.*, 2016)
CYP19A1	Prostatic Breast	rs700519 rs10046 rs1008805 rs1062033 rs2289105 rs4646 rs4775936 rs749292 rs6493497 rs7176005	Missense *19 C>T Substitution Substitution Substitution *161T>G Substitution Substitution Substitution Substitution	Arg264Cys 3' UTR Intronic Intronic Intronic 3'UTR Intronic Intronic 5'Flanking 5'Flanking	Asians Mixed Mixed European	antiandrogens hormone Therapy hmg coa reductase inhibitors hdl cholesterol triglycerides letrozole enzyme inhibitors	(Yu *et al.*, 2013) (Santa-Maria *et al.*, 2016) (Oesterreich *et al.*, 2015) (Fontein *et al.*, 2014)

GSTP1	Lung Uterine Cervical Breast	rs1695 rs1695 rs1695	Missense Missense Missense	Ile105Val Ile105Val Ile105Val	Mixed Russian	oxaliplatin fluorouracil leucovorin cisplatin radiotherapy cisplatin docetaxel cyclophosphamide fluorouracil methotrexate doxorubicin epirubicin	(Booton *et al.*, 2006) (Khrunin *et al.*, 2012) (Zhang *et al.*, 2011)
CYP1B1	Prostatic Breast	rs1056836 rs1056836	Missense Missense	Leu432Val Leu432Val	European Asian	docetaxel paclitaxel gemcitabine	(Sissung *et al.*, 2008) (Lee *et al.*, 2014)
CYP3A4	Breast	*1 *8 *20 *22 rs2740574	Multiple Alleles Substitution	 ‘5 Flanking	European/Asian	paclitaxel fluorouracil doxorubicin docetaxel cyclophosphamide methotrexate epirubicin paclitaxel tamoxifen	(Su *et al.*, 2010) (Tang *et al.*, 2013) (Boso *et al.*, 2014) (Chu *et al.*, 2007)
CYP3A5	Breast	rs776746 *1A *3A	Substitution Multiple Alleles	Intronic	Asian	paclitaxel docetaxel, gemcitabine tamoxifen	(Tang *et al.*, 2013) (Saladores *et al.*, 2015)

ESR1	Breast	rs2813543 rs4870061 rs9322335 rs9322336 rs9340799	Substitution Substitution Substitution Substitution Substitution	3' Flanking Intronic Intronic Intronic Intronic	Mixed Mixed Whites Asian	exemestane letrozole, tamoxifen	(Oesterreich *et al.*, 2015) (Henry *et al.*, 2013) (Ntukidem *et al.*, 2008) (Wang *et al.*, 2013)
CYP2D6	Breast	*1 *7 *1xN *9 *2 *10 *2A *11 *2xN *14 *3 *17 *4 *10x2 *4A *20 *4D *21 *5 *29 *6 *35 *6B *36	Multiple Alleles		African European Asian	tamoxifen 4-hydroxytamoxifen n-desmethyltamoxifen anthracyclines antineoplastic agents endoxifen enzyme inhibitors leuprolide raloxifene	(Wright *et al.*, 2010) (Kurose *et al.*, 2012) (Lyon *et al.*, 2012) (de Vries Schultink *et al.*, 2015)
XRCC1	Uterine/Cervical	rs25487	Missense	Gln399Arg	Asian	oxaliplatin, fluorouracil cisplatin, carboplatin	(Chung *et al.*, 2006) (Lee *et al.*, 2013)
	Oesophageal	rs25487	Missense	Gln399Arg	Multiple	leucovorin oxaliplatin fluorouracil	(Findlay *et al.*, 2015)
XRCC5	Lung	rs25487 rs6941	Missense 3' UTR	Gln399Arg	Asian Asian	Platinum	(Kalikaki *et al.*, 2009) (Chen *et al.*, 2016)
	Multiple Myeloma	rs1051685	3' UTR		European	thalidomide	(Cibeira *et al.*, 2011)
SLC19A1	Uterine Cervical	rs12659	Synonymous	Pro192Pro		cisplatin fluorouracil	(Chung *et al.*, 2006)

	Lung	rs1051298	3' UTR			carboplatin bevacizumab pemetrexed	(Corrigan *et al.*, 2014)

*Drugs are listed in summary of all SNPs involved in the drug-gene interactions highlighted in the cancer type.

Note on Star Alleles: Star alleles are haplotypes of Cytochrome P450 (CYP) genes characterised by a star and a numeral in order to shorten their nomenclature. This is done as multiple SNPs, indels or copy number variants may be forming the association signal for a drug adverse reaction. A full list of star allele nomenclature can be accessed at: http://www.cypalleles.ki.se/ ;

CYP3A4: http://www.cypalleles.ki.se/cyp3a4.htm

CYP3A5: http://www.cypalleles.ki.se/cyp3a5.htm

CYP2D6: http://www.cypalleles.ki.se/cyp2d6.htm

Reference List for Appendix Table 2

Booton, R., Ward, T., Heighway, J., Ashcroft, L., Morris, J., and Thatcher, N. (2006). Glutathione-S-transferase P1 isoenzyme polymorphisms, platinum-based chemotherapy, and non-small cell lung cancer. J. Thorac. Oncol. 1, 679–683.

Boso, V., Herrero, M.J., Santaballa, A., Palomar, L., Megias, J.E., de la Cueva, H., Rojas, L., Marques, M.R., Poveda, J.L., Montalar, J., et al. (2014). SNPs and taxane toxicity in breast cancer patients. Pharmacogenomics 15, 1845–1858.

Chen, J., Wu, L., Wang, Y., Yin, J., Li, X., Wang, Z., Li, H., Zou, T., Qian, C., Li, C., et al. (2016). Effect of transporter and DNA repair gene polymorphisms to lung cancer chemotherapy toxicity. Tumour Biol. 37, 2275–2284.

Chu, W., Fyles, A., Sellers, E.M., McCready, D.R., Murphy, J., Pal, T., and Narod, S.A. (2007). Association between CYP3A4 genotype and risk of endometrial cancer following tamoxifen use. Carcinogenesis 28, 2139–2142.

Chung, H.H., Kim, M.-K., Kim, J.W., Park, N.-H., Song, Y.-S., Kang, S.-B., and Lee, H.-P. (2006). XRCC1 R399Q polymorphism is associated with response to platinum-based neoadjuvant chemotherapy in bulky cervical cancer. Gynecol. Oncol. 103, 1031–1037.

Cibeira, M.T., de Larrea, C.F., Navarro, A., Diaz, T., Fuster, D., Tovar, N., Rosinol, L., Monzo, M., and Blade, J. (2011). Impact on response and survival of DNA repair single nucleotide polymorphisms in relapsed or refractory multiple myeloma patients treated with thalidomide. Leuk. Res. 35, 1178–1183.

Corrigan, A., Walker, J.L., Wickramasinghe, S., Hernandez, M.A., Newhouse, S.J., Folarin, A.A., Lewis, C.M., Sanderson, J.D., Spicer, J., and Marinaki, A.M. (2014). Pharmacogenetics of pemetrexed combination therapy in lung cancer: pathway analysis reveals novel toxicity associations. Pharmacogenomics J. 14, 411–417.

Findlay, J.M., Middleton, M.R., and Tomlinson, I. (2015). A systematic review and meta-analysis of somatic and germline DNA sequence biomarkers of esophageal cancer survival, therapy response and stage. Ann. Oncol. 26, 624–644.

Fontein, D.B.Y., Houtsma, D., Nortier, J.W.R., Baak-Pablo, R.F., Kranenbarg, E.M.-K., van der Straaten, T.R.J.H.M., Putter, H., Seynaeve, C., Gelderblom, H., van de Velde, C.J.H., et al. (2014). Germline variants in the CYP19A1 gene are related to specific adverse events in aromatase inhibitor users: a substudy of Dutch patients in the TEAM trial. Breast Cancer Res. Treat. 144, 599–606.

Goff, L.W., Thakkar, N., Du, L., Chan, E., Tan, B.R., Cardin, D.B., McLeod, H.L., Berlin, J.D., Zehnbauer, B., Fournier, C., et al. (2014). Thymidylate synthase genotype-directed chemotherapy for patients with gastric and gastroesophageal junction cancers. PLoS One 9, e107424.

Gross, E., Ullrich, T., Seck, K., Mueller, V., de Wit, M., von Schilling, C., Meindl, A., Schmitt, M., and Kiechle, M. (2003). Detailed analysis of five mutations in dihydropyrimidine dehydrogenase detected in cancer patients with 5-fluorouracil-related side effects. Hum. Mutat. 22, 498.

Henry, N.L., Skaar, T.C., Dantzer, J., Li, L., Kidwell, K., Gersch, C., Nguyen, A.T., Rae, J.M., Desta, Z., Oesterreich, S., et al. (2013). Genetic associations with toxicity-related discontinuation of aromatase inhibitor therapy for breast cancer. Breast Cancer Res. Treat. 138, 807–816.

Kalikaki, A., Kanaki, M., Vassalou, H., Souglakos, J., Voutsina, A., Georgoulias, V., and Mavroudis, D. (2009). DNA repair gene polymorphisms predict favorable clinical outcome in advanced non-small-cell lung cancer. Clin. Lung Cancer 10, 118–123.

Khrunin, A., Ivanova, F., Moisseev, A., Khokhrin, D., Sleptsova, Y., Gorbunova, V., and Limborska, S. (2012). Pharmacogenomics of cisplatin-based chemotherapy in ovarian cancer patients of different ethnic origins. Pharmacogenomics 13, 171–178.

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

Lal, S., Wong, Z.W., Sandanaraj, E., Xiang, X., Ang, P.C.S., Lee, E.J.D., and Chowbay, B. (2008). Influence of ABCB1 and ABCG2 polymorphisms on doxorubicin disposition in Asian breast cancer patients. Cancer Sci. 99, 816–823.

Lee, K.-H., Chang, H.J., Han, S.-W., Oh, D.-Y., Im, S.-A., Bang, Y.-J., Kim, S.Y., Lee, K.-W., Kim, J.H., Hong, Y.S., et al. (2013). Pharmacogenetic analysis of adjuvant FOLFOX for Korean patients with colon cancer. Cancer Chemother. Pharmacol. 71, 843–851.

Lee, S.-Y., Im, S.-A., Park, Y.H., Woo, S.Y., Kim, S., Choi, M.K., Chang, W., Ahn, J.S., and Im, Y.-H. (2014). Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. Eur. J. Cancer 50, 698–705.

Lyon, E., Gastier Foster, J., Palomaki, G.E., Pratt, V.M., Reynolds, K., Sábato, M.F., Scott, S. a, and Vitazka, P. (2012). Laboratory testing of CYP2D6 alleles in relation to tamoxifen therapy. Genet. Med. 14, 990–1000.

Megias-Vericat, J.E., Rojas, L., Herrero, M.J., Boso, V., Montesinos, P., Moscardo, F., Poveda, J.L., Sanz, M.A., and Alino, S.F. (2015). Influence of ABCB1 polymorphisms upon the effectiveness of standard treatment for acute myeloid leukemia: a systematic review and meta-analysis of observational studies. Pharmacogenomics J. 15, 109–118.

Meulendijks, D., Jacobs, B.A.W., Aliev, A., Pluim, D., van Werkhoven, E., Deenen, M.J., Beijnen, J.H., Cats, A., and Schellens, J.H.M. (2016). Increased risk of severe fluoropyrimidine-associated toxicity in patients carrying a G to C substitution in the first 28-bp tandem repeat of the thymidylate synthase 2R allele. Int. J. Cancer 138, 245–253.

Narumiya, K., Metzger, R., Bollschweiler, E., Alakus, H., Brabender, J., Drebber, U., Holscher, A.H., and Warnecke-Eberz, U. (2011). Impact of ABCB1 C3435T polymorphism on lymph node regression in multimodality treatment of locally advanced esophageal cancer. Pharmacogenomics 12, 205–214.

Ntukidem, N.I., Nguyen, A.T., Stearns, V., Rehman, M., Schott, A., Skaar, T., Jin, Y., Blanche, P., Li, L., Lemler, S., et al. (2008). Estrogen receptor genotypes, menopausal status, and the lipid effects of tamoxifen. Clin. Pharmacol. Ther. 83, 702–710.

Oesterreich, S., Henry, N.L., Kidwell, K.M., Van Poznak, C.H., Skaar, T.C., Dantzer, J., Li, L., Hangartner, T.N., Peacock, M., Nguyen, A.T., et al. (2015). Associations between genetic variants and the effect of letrozole and exemestane on bone mass and bone turnover. Breast Cancer Res. Treat. 154, 263–273.

Saladores, P., Murdter, T., Eccles, D., Chowbay, B., Zgheib, N.K., Winter, S., Ganchev, B., Eccles, B., Gerty, S., Tfayli, A., et al. (2015). Tamoxifen metabolism predicts drug concentrations and outcome in premenopausal patients with early breast cancer. Pharmacogenomics J. 15, 84–94.

Santa-Maria, C.A., Blackford, A., Nguyen, A.T., Skaar, T.C., Philips, S., Oesterreich, S., Rae, J.M., Desta, Z., Robarge, J., Henry, N.L., et al. (2016). Association of Variants in Candidate Genes with Lipid Profiles in Women with Early Breast Cancer on Adjuvant Aromatase Inhibitor Therapy. Clin. Cancer Res. 22, 1395–1402.

Sissung, T.M., Danesi, R., Price, D.K., Steinberg, S.M., de Wit, R., Zahid, M., Gaikwad, N., Cavalieri, E., Dahut, W.L., Sackett, D.L., et al. (2008). Association of the CYP1B1*3 allele with survival in patients with prostate cancer receiving docetaxel. Mol. Cancer Ther. 7, 19–26.

Su, H.I., Sammel, M.D., Velders, L., Horn, M., Stankiewicz, C., Matro, J., Gracia, C.R., Green, J., and DeMichele, A. (2010). Association of cyclophosphamide drug-metabolizing enzyme polymorphisms and chemotherapy-related ovarian failure in breast cancer survivors. Fertil. Steril. 94, 645–654.

Tang, N.L.S., Liao, C. Di, Wang, X., Mo, F.K.F., Chan, V.T.C., Ng, R., Pang, E., Suen, J.J.S., Woo, J., and Yeo, W. (2013). Role of pharmacogenetics on adjuvant chemotherapy-induced neutropenia in Chinese breast cancer patients. J. Cancer Res. Clin. Oncol. 139, 419–427.

Toffoli, G., Giodini, L., Buonadonna, A., Berretta, M., De Paoli, A., Scalone, S., Miolo, G., Mini, E., Nobili, S., Lonardi, S., et al. (2015). Clinical validity of a DPYD-based pharmacogenetic test to predict severe toxicity to fluoropyrimidines. Int. J. Cancer 137, 2971–2980.

de Vries Schultink, A.H.M., Zwart, W., Linn, S.C., Beijnen, J.H., and Huitema, A.D.R. (2015). Effects of Pharmacogenetics on the Pharmacokinetics and Pharmacodynamics of Tamoxifen. Clin. Pharmacokinet. 54, 797–810.

Wang, J., Lu, K., Song, Y., Xie, L., Zhao, S., Wang, Y., Sun, W., Liu, L., Zhao, H., Tang, D., et al. (2013). Indications of clinical and genetic predictors for aromatase inhibitors related musculoskeletal adverse events in Chinese Han women with breast cancer. PLoS One 8, e68798.

Wright, G.E.B., Niehaus, D.J.H., Drogemuller, B.I., Koen, L., Gaedigk, A., and Warnich, L. (2010). Elucidation of CYP2D6 genetic diversity in a unique African population: Implications for the future application of pharmacogenetics in the xhosa population. Ann. Hum. Genet. 74, 340–350.

Yu, C.-C., Huang, S.-P., Lee, Y.-C., Huang, C.-Y., Liu, C.-C., Hour, T.-C., Huang, C.-N., You, B.-J., Chang, T.-Y., Huang, C.-H., et al. (2013). Molecular markers in sex hormone pathway genes associated with the efficacy of androgen-deprivation therapy for prostate cancer. PLoS One 8, e54627.

Zhang, B.-L., Sun, T., Zhang, B.-N., Zheng, S., Lu, N., Xu, B.-H., Wang, X., Chen, G.-J., Yu, D.-K., and Lin, D.-X. (2011). Polymorphisms of GSTP1 is associated with differences of chemotherapy response and toxicity in breast cancer. Chin. Med. J. (Engl). 124, 199–204.

Zhao, X.-Q., Cao, W.-J., Yang, H.-P., Yang, X.-W., Tang, P., Sun, L., and Gao, X. (2016). DPYD gene polymorphisms are associated with risk and chemotherapy prognosis in pediatric patients with acute lymphoblastic leukemia. Tumour Biol.

Annotation and Data Scripts
The following scripts were used for various stages of data analysis and variant annotation

7.2.1 Variant Effect Predictor Running Script

```
# enable variant effect predictor

./vep \

# population specific VCF file as input, contains all variants in selected PGx genes

-i  pgx_genes_pop.vcf \

# connect to Ensembl's cache base of models, regulatory features and variant data for a species

--cache \

# use the cache for homo sapiens GRCh37

--port 3337 \

# informative output filename

-o pop_vep_output.txt \

# output results in tab separated delimiters, with even spacing between columns

--tab \

# run Polyphen-2 with both prediction and score readings

--polyphen b \

# run SIFT with both prediction and score readings

--sift b \

# label variants whose predicted effects occur in the canonical transcript

--canonical \

# attach gene symbol to variant

--symbol \

# retrieve allele frequencies for KGP combined populations

--af_1kg \

# activate CAROL analysis

--plugin Carol \
```

activate Condel analysis

--plugin Condel,/home/jorge/.vep/config/Condel/config/,b,2 \

activate CADD with selected datafile as reference

--plugin CADD,whole_genome_SNVs.tsv.gz,1000G_phase3.tsv.gz \

activate FATHMM-MKL with selected datafile

--plugin FATHMM_MKL,fathmm-MKL_Current.tab.gz

activate LOFTEE with selected datafile

--plugin loftee, LoF,human_ancestor_fa

7.2.2 Plink v1.9 Frequency Determination

activate Plink

./Plinkv1.9 \

specify vcf file to use as input

--vcf pgx_genes_pop.vcf \

#call frequencies informatively and report on homozygote and heterozygote numbers

--freqx \

output frequency calls to file

--out pop_freq

7.2.3 Sorting

In order to merge the VEP outputs and the frequencies called by Plink, both filesets must be sorted:

```
#initiate sorting tool

./sort \

# sort file by column 2, which specifies to sort by variant location

-k2,2 \

# input files to be sorted, both vepouts and frequencies must be sorted

pop_vep_output.txt   \

#output to file

> pop_vep_output_sorted.txt
```

7.2.4 Merging

```
#initiate joining tool

./join \

# merge column 1 of the sorted vep output file to the sorted frequency call file

-1 1 -2 2 \

# indicate to use sorted vep output file

pop_vep_output_sorted.txt   \

# indicate the sorted frequency call file

pop_freqs_sorted.txt \

#output to file

> pop_joined_vep_freqs.txt
```

Statistical Analysis Scripts

The file to be used as input (allele_fileset) consists of the headers Gene; ID; Pop1; Pop2; … Pop8 , and two rows per variant, the first row being the count of the minor allele in the respective Pop, the second row being the count of the remaining alleles for that variant.

e.g. allele_fileset

Gene	ID	YRI	ESN	GWD	MSL	LWK	ZULU	BAG	ETHIO
ABCB1	rs1045642	27	23	43	25	28	19	19	37
ABCB1	rs1045642	189	175	183	145	170	181	181	203

7.2.5 Single Statistical Test

This was used to assess if difference in alleles between population sets is significant, which was used for only African populations.

```
# define a variable “b” which will store processed data, a matrix transposed by t()

# sapply() is used to loop the data outputs of the statistical test used into simplified data matrices

# seq() feeds two rows into the function command, making inputs of two single lines (for both alleles for a given variant), and stores these in a variable called “d1”

# “d1” is then used as the input for a statistical test, the results of which are stored in the variable “ch”

# the columns on which to apply these are defined in 3:9, which can be adjusted for use in global population comparisons

b<-t(sapply(seq(1,nrow(allele_fileset),2),function(x,Data){d1<-Data[x:(x+1),3:9];

# The statistical test here, is a Chi-Square test, given by chisq.test(), a fisher test can also be used here, via fisher.test() (example given in next script).

# when the fisher test is used, a workspace of memory must be defined, e.g. fisher.test(x, workspace = 200000);

# return() returns a data-frame of the Gene and ID, along with the p-values stored in “ch”

ch<-chisq.test(d1);return(data.frame(Data[x,1:2],ch$p.value))},allele_fileset))

#convert the data in “b” as a data-frame named for convenience

Stats_results <- as.data.frame(b)
```

7.2.6 Pairwise Statistical Test Script

This was used to compare individual African populations each other, and Africans and Global populations (example file below) to assess if differences in allele frequencies are significant.

SYMBOL	ID	Con_AFR	AMR_AF	EAS_AF	EUR_AF	SAS_AF
ABCB1	rs1045642	221	297	401	521	562
ABCB1	rs1045642	1427	397	607	485	416

```
# define a variable “NamePop” which stores the population names in the input file

NamePop<-names(allele_fileset)

#defines a variable to build a loop with if() and else in later steps

Cmt=1

# define the columns for which to run the test, numbers can be adjusted to suit the input

for(col1 in 3:7){

  for(col2 in (col1+1):8){

#   Similar to the function above, this specifies to run the test between the columns specified above.

 a<-t(sapply(seq(1,nrow(African_vars_stats),2),function(x,Data, col1, col2){d1<-Data[x:(x+1),c(col1,col2)];ch<-fisher.test(d1, workspace = 200000);return(c(Data[x,1:2],ch$p.value))},African_vars_stats,col1,col2))

#   This loop defines data to be stored into a data-frame, which makes use of rbind() to stitch the data into a neat matrix

 a<-as.data.frame(a)

    a$Pop1<-NamePop[col1]

    a$Pop2<-NamePop[col2]

    if(Cmt==1)Stats_results<-a

    else Stats_results<-rbind(Stats_results,a)

    Cmt=Cmt+1

  }

}
```

Appendix Table 3: Summary of variant counts in each evaluated gene per population analysed

Bagandan (BAG)	***ABCB1***	***DPYD***	***TYMS***	***CYP19A1***	***GSTP1***	***CYP1B1***	***CYP3A4***	***CYP3A5***	***ESR1***	***CYP2D6***	***XRCC1***	***XRCC5***
3 Prime UTR Variant	9	6	7	18	5	25	0	1	24	0	1	6
5 Prime UTR Variant	2	0	1	1	0	2	0	1	2	1	2	5
Downstream Gene Variant	2	5	10	3	6	7	8	9	26	11	3	8
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	0	0	0
Intron Variant	1296	5531	132	993	21	26	118	133	3217	29	212	698
Missense Variant	6	10	2	2	3	7	4	3	3	10	9	1
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	2	2	0	1	0	0	1	1	0	0	2	1
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	0	0	0	0	0	0	0	0	0	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	6	6	2	5	2	7	1	2	7	5	6	2
Upstream Gene Variant	11	8	19	10	14	46	6	7	49	5	34	4
BAG Total	1334	5568	173	1033	51	120	138	158	3328	62	269	725
Esan (ESN)	***ABCB1***	***DPYD***	***TYMS***	***CYP19A1***	***GSTP1***	***CYP1B1***	***CYP3A4***	***CYP3A5***	***ESR1***	***CYP2D6***	***XRCC1***	***XRCC5***
3 Prime UTR Variant	8	4	7	15	1	27	1	2	26	0	3	9
5 Prime UTR Variant	2	0	0	2	2	2	0	1	2	3	3	6
Downstream Gene Variant	3	4	13	3	6	7	15	10	33	10	3	8
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	1	0	0
Intron Variant	1311	5468	142	947	21	26	148	152	2591	40	257	715
Missense Variant	8	9	1	3	4	6	4	1	3	14	9	2
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	2	1	0	0	0	0	1	1	1	1	3	0
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	1	0	0	0	0	0	0	0	0	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	4	5	0	4	2	5	1	2	7	6	5	3
Upstream Gene Variant	7	10	21	7	16	49	5	11	61	9	42	5
ESN Total	1345	5493	184	981	52	122	175	181	2724	85	325	748

Ethiopian (ETHIO)	*ABCB1*	*DPYD*	*TYMS*	*CYP19A1*	*GSTP1*	*CYP1B1*	*CYP3A4*	*CYP3A5*	*ESR1*	*CYP2D6*	*XRCC1*	*XRCC5*
3 Prime UTR Variant	8	5	5	19	0	26	1	1	23	0	1	11
5 Prime UTR Variant	2	1	1	1	1	1	2	3	3	2	2	7
Downstream Gene Variant	3	8	10	3	6	7	5	8	20	13	3	9
Frameshift Variant	0	0	0	0	0	0	0	2	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	0	0	0
Intron Variant	1174	5444	130	957	23	26	134	150	2978	32	237	707
Missense Variant	6	9	1	4	4	7	3	6	3	11	11	2
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	2	2	0	0	0	0	0	2	0	0	4	2
Start Lost	0	0	0	0	0	0	0	0	1	0	0	0
Stop Gained	1	0	0	0	0	0	0	1	0	0	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	5	7	3	6	1	7	2	2	8	7	4	3
Upstream Gene Variant	6	4	18	11	17	41	8	7	47	6	34	5
ETHIO Total	1207	5480	168	1001	69	115	155	182	3083	72	296	746
Gambian (GWD)	*ABCB1*	*DPYD*	*TYMS*	*CYP19A1*	*GSTP1*	*CYP1B1*	*CYP3A4*	*CYP3A5*	*ESR1*	*CYP2D6*	*XRCC1*	*XRCC5*
3 Prime UTR Variant	9	3	7	14	0	29	0	1	26	0	3	8
5 Prime UTR Variant	2	0	1	2	3	1	1	1	1	1	2	6
Downstream Gene Variant	4	4	15	3	6	9	16	8	35	14	4	7
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	1	0	0
Intron Variant	1344	5814	156	926	26	23	135	175	3162	43	239	779
Missense Variant	10	12	0	3	1	5	2	0	1	16	8	1
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	2	1	0	0	0	0	0	1	1	1	1	1
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	0	0	0	0	0	0	1	0	1	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	6	4	0	3	3	7	3	2	8	6	3	3
Upstream Gene Variant	6	9	22	7	16	51	4	9	49	12	39	10
GWD Total	1383	5847	201	958	55	125	161	199	3283	96	299	815

Luhya (LWK)	***ABCB1***	***DPYD***	***TYMS***	***CYP19A1***	***GSTP1***	***CYP1B1***	***CYP3A4***	***CYP3A5***	***ESR1***	***CYP2D6***	***XRCC1***	***XRCC5***
3 Prime UTR Variant	9	5	6	13	3	27	1	1	23	0	3	9
5 Prime UTR Variant	2	0	0	2	0	1	0	2	2	2	3	6
Downstream Gene Variant	3	3	14	3	9	8	11	9	26	14	3	7
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	0	0	0
Intron Variant	1352	5791	133	994	22	30	129	159	3149	38	243	701
Missense Variant	5	8	1	2	2	8	4	1	1	14	10	1
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	2	1	0	1	0	0	0	1	0	2	2	3
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	0	0	0	0	0	0	1	0	0	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	6	4	2	5	2	8	2	3	7	9	3	2
Upstream Gene Variant	5	10	20	8	18	49	7	6	52	13	43	7
LWK Total	1384	5822	176	1028	56	131	154	184	3260	93	310	736
Mende (MSL)	***ABCB1***	***DPYD***	***TYMS***	***CYP19A1***	***GSTP1***	***CYP1B1***	***CYP3A4***	***CYP3A5***	***ESR1***	***CYP2D6***	***XRCC1***	***XRCC5***
3 Prime UTR Variant	8	5	7	13	0	27	2	1	24	0	3	8
5 Prime UTR Variant	1	0	1	1	2	1	0	1	0	1	2	6
Downstream Gene Variant	3	6	15	3	4	8	17	9	39	15	4	9
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	1	0	0
Intron Variant	1340	5330	171	908	24	29	128	156	3087	41	236	704
Missense Variant	9	9	0	2	3	6	5	0	2	12	8	1
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	1	2	0	0	0	0	0	1	0	1	2	0
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	0	0	0	0	0	0	0	0	1	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	7	4	1	4	3	7	1	3	6	6	4	1
Upstream Gene Variant	8	11	23	11	15	46	3	8	58	8	47	8
MSL Total	1377	5367	218	942	51	124	156	180	3216	87	306	737

Yoruba (YRI)	*ABCB1*	*DPYD*	*TYMS*	*CYP19A1*	*GSTP1*	*CYP1B1*	*CYP3A4*	*CYP3A5*	*ESR1*	*CYP2D6*	*XRCC1*	*XRCC5*
3 Prime UTR Variant	9	4	7	18	0	27	1	1	27	0	3	6
5 Prime UTR Variant	2	2	0	2	2	2	0	1	1	3	2	6
Downstream Gene Variant	3	6	14	5	5	6	14	9	33	13	3	9
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	1	0	0
Intron Variant	1353	5577	154	920	20	28	147	161	3214	37	238	720
Missense Variant	7	13	0	2	1	6	2	0	4	14	7	1
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	2	2	1	1	0	0	1	1	0	0	2	1
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	0	0	0	0	0	0	0	0	0	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	6	4	0	6	2	8	1	2	7	6	5	1
Upstream Gene Variant	7	7	24	7	15	49	3	10	61	9	46	4
YRI Total	1389	5615	200	961	45	126	169	186	3347	84	306	748
Zulu (ZULU)	**ABCB1**	**DPYD**	**TYMS**	**CYP19A1**	**GSTP1**	**CYP1B1**	**CYP3A4**	**CYP3A5**	**ESR1**	**CYP2D6**	**XRCC1**	**XRCC5**
3 Prime UTR Variant	8	5	7	22	0	27	0	1	21	1	1	9
5 Prime UTR Variant	2	1	1	1	2	1	0	3	1	0	2	4
Downstream Gene Variant	2	5	11	5	8	9	10	9	23	11	1	7
Frameshift Variant	0	0	0	0	0	0	0	1	0	0	0	0
In-frame Deletion	0	0	0	0	0	0	0	0	0	0	0	0
In-frame Insertion	0	0	0	0	0	0	0	0	0	0	0	0
Intron Variant	1169	5531	135	971	20	25	104	133	2882	27	228	688
Missense Variant	4	12	2	2	1	7	5	3	1	11	10	1
Splice Acceptor Variant	0	0	0	0	0	0	0	0	0	1	0	0
Splice Donor Variant	0	0	0	0	0	0	0	0	0	0	0	0
Splice Region Variant, Intron Variant	3	3	0	1	0	0	0	1	0	0	1	0
Start Lost	0	0	0	0	0	0	0	0	0	0	0	0
Stop Gained	0	0	0	0	0	0	0	0	0	0	0	0
Stop Lost	0	0	0	0	0	0	0	0	0	0	0	0
Synonymous Variant	4	4	0	5	3	7	2	3	5	6	4	2
Upstream Gene Variant	7	6	18	10	16	46	8	7	53	4	36	3
ZULU Total	1199	5567	174	1017	50	122	129	161	2986	61	303	714

Appendix B Table 4: Variant population frequency data for continental Africans and KGP global populations

SYMBOL	ID	Location	Allele	Consequence	Population Allele Frequency				
					Cont AFR	AMR	EAS	EUR	SAS
ABCB1	7:87195405	7:87195405	T	Stop Gained	0.0006	0	0	0	0
ABCB1	rs1045642	7:87138645	A	Synonymous Variant	0.134	0.428	0.398	0.518	0.575
ABCB1	rs1128503	7:87179601	A	Synonymous Variant	0.128	0.404	0.627	0.416	0.587
ABCB1	rs137996914	7:87133567	A	Missense Variant	0.0006	0	0.001	0	0
ABCB1	rs142600685	7:87179256	A	Missense Variant	0.0006	0	0	0	0
ABCB1	rs147487745	7:87150189	T	Missense Variant	0.0006	0	0	0	0
ABCB1	rs148718120	7:87150095	G	Missense Variant	0.0024	0	0	0	0
ABCB1	rs200754866	7:87165800	T	Missense Variant	0.0006	0	0	0	0
ABCB1	rs201159898	7:87178825	A	Missense Variant	0.0006	0	0	0	0
ABCB1	rs2032582	7:87160618	A	Missense Variant	0.0212	0.369	0.398	0.41	0.592
ABCB1	rs2032583	7:87160561	G	Intron Variant	0.212	0.112	0.069	0.134	0.18
ABCB1	rs2229109	7:87179809	T	Missense Variant	0.0042	0.025	0	0.033	0.009
ABCB1	rs35730308	7:87138758	G	Missense Variant	0.0018	0	0	0	0
ABCB1	rs535338471	7:87179198	C	Missense Variant	0.0006	0	0	0	0
ABCB1	rs572038993	7:87145971	G	Missense Variant	0.0006	0	0	0	0
ABCB1	rs930790323	7:87179284	G	Missense Variant	0.0006	0	0	0	0
CYP19A1	rs1008805	15:51549599	G	Intron Variant	0.160	0.566	0.210	0.447	0.444
CYP19A1	rs142652579	15:51514707	T	Missense Variant	0.0006	0	0	0	0
CYP19A1	rs143562020	15:51510760	T	Missense Variant	0.0006	0	0	0	0
CYP19A1	rs4646	15:51502844	A	3 Prime UTR Variant	0.298	0.473	0.297	0.290	0.395
CYP19A1	rs6493497	15:51630835	A	Upstream Gene Variant	0.299	0.184	0.200	0.149	0.272
CYP19A1	rs749292	15:51558731	A	Intron Variant	0.430	0.339	0.493	0.401	0.313
CYP1B1	rs1056836	2:38298203	G	Missense	0.208	0.723	0.909	0.602	0.831
CYP1B1	rs1800440	2:38298139	C	Missense Variant	0.007	0.131	0.005	0.196	0.201
CYP1B1	rs747324108	2:38298091	T	Missense Variant	0.0006	0	0	0	0
CYP1B1	rs9282671	2:38302291	T	Missense Variant	0.002	0.001	0	0.007	0.002
CYP2D6	rs1058172	22:42523528	T	Missense Variant	0.006	0	0	0	0
CYP2D6	rs1065852	22:42526694	A	Missense Variant	0.089	0.148	0.571	0.202	0.165
CYP2D6	rs143145507	22:42523891	A	Missense Variant	0.001	0	0	0	0
CYP2D6	rs147960066	22:42523592	A	Stop Gained	0.0006	0	0	0	0
CYP2D6	rs16947	22:42523943	G	Missense Variant	0.405	0.673	0.860	0.657	0.638
CYP2D6	rs28371703	22:42525821	T	Missense Variant	0.012	0.095	0.001	0.173	0.081
CYP2D6	rs369177208	22:42522721	T	Missense Variant	0.002	0	0	0	0

CYP2D6	rs373243894	22:42526672	A	Missense Variant	0.0006	0	0	0	0
CYP2D6	rs376636053	22:42525158	G	Missense Variant	0.0006	0.001	0	0	0
CYP2D6	rs3892097	22:42524947	T	Splice Acceptor Variant	0.043	0.130	0.002	0.186	0.109
CYP2D6	rs532668079	22:42522748	T	Missense Variant	0.0006	0	0	0	0
CYP2D6	rs536109057	22:42526640	A	Stop Gained	0.001	0	0	0	0
CYP2D6	rs59421388	22:42523610	T	Missense Variant	0.117	0.003	0	0	0
CYP3A4	7:99375691	7:99375691	T	Missense Variant	0.0006	0	0	0	0
CYP3A4	rs12721629	7:99359800	A	Missense Variant	0.007	0.001	0	0	0
CYP3A4	rs2242480	7:99361466	C	Intron Variant	0.183	0.607	0.732	0.919	0.629
CYP3A4	rs2740574	7:99382096	T	Upstream Gene Variant	0.257	0.895	0.996	0.972	0.960
CYP3A4	rs35599367	7:99366316	A	Intron Variant	0.002	0.026	0	0.050	0.006
CYP3A4	rs4646440	7:99360870	A	Intron Variant	0.077	0.236	0.216	0.003	0.263
CYP3A4	rs530166880	7:99358480	A	Missense Variant	0.0006	0	0	0	0
CYP3A4	rs540702483	7:99361630	T	Missense Variant	0.0006	0	0	0	0
CYP3A4	rs57409622	7:99367428	A	Missense Variant	0.002	0	0	0	0
CYP3A4	rs67784355	7:99359829	A	Missense Variant	0.0006	0	0	0.001	0
CYP3A4	rs897005504	7:99377646	T	Missense Variant	0.002	0	0	0	0
CYP3A5	7:99270249	7:99270249	G	Missense Variant	0.002	0	0	0	0
CYP3A5	7:99270300	7:99270300	C	Missense Variant, Splice Region Variant	0.0006	0	0	0	0
CYP3A5	rs111371159	7:99260462	T	Stop Gained	0.0006	0	0	0	0
CYP3A5	rs142823108	7:99264299	G	Missense Variant	0.001	0	0	0	0
CYP3A5	rs145774441	7:99260477	G	Missense Variant	0.0006	0	0.002	0	0.001
CYP3A5	rs149664815	7:99247731	A	Stop Gained	0.002	0	0	0	0
CYP3A5	rs188366390	7:99258153	G	Missense Variant	0.007	0	0	0.001	0
CYP3A5	rs200579169	7:99273810-99273811	C	Frameshift Variant	0.0006	0.001	0	0.012	0.004
CYP3A5	rs41303343	7:99250393-99250394	A	Frameshift Variant	0.098	0.003	0	0	0
CYP3A5	rs776746	7:99270539	C	Intron Variant	0.195	0.797	0.713	0.943	0.668
DPYD	COSM1688098	1:97847978	A	Missense Variant	0.0006	0	0	0	0
DPYD	rs114096998	1:97544543	T	Missense Variant	0.023	0	0	0	0
DPYD	rs115232898	1:98165030	C	Missense Variant	0.022	0.003	0	0	0
DPYD	rs12022243	1:97862780	T	Intron Variant	0.163	0.137	0.220	0.199	0.406
DPYD	rs142619737	1:97981407	T	Missense Variant	0.002	0	0	0	0
DPYD	rs144395748	1:98015282	C	Missense Variant	0.001	0	0	0	0
DPYD	rs145548112	1:97771751	T	Missense Variant	0.0006	0.001	0	0	0
DPYD	rs17376848	1:97915624	G	Synonymous Variant	0.007	0.082	0.114	0.037	0.035
DPYD	rs1801158	1:97981421	T	Missense Variant	0.0006	0.013	0	0.031	0.005
DPYD	rs1801159	1:97981395	C	Missense Variant	0.170	0.267	0.266	0.188	0.083
DPYD	rs1801160	1:97770920	T	Missense Variant	0.032	0.055	0.015	0.050	0.087
DPYD	rs1801265	1:98348885	G	Missense Variant	0.427	0.225	0.086	0.215	0.266
DPYD	rs201268750	1:97544676	T	Missense Variant	0.0006	0	0	0	0
DPYD	rs2297595	1:98165091	C	Missense Variant	0.095	0.063	0.018	0.119	0.061

DPYD	rs573299212	1:98058829	T	Missense Variant	0.001	0	0	0	0.001
DPYD	rs60511679	1:97770919	C	Missense Variant	0.001	0	0	0	0
DPYD	rs61622928	1:98039437	T	Missense Variant	0.090	0.004	0	0	0
DPYD	rs72549310	1:98348909	A	Stop Gained	0.0006	0	0	0	0
DPYD	rs72975710	1:98015291	A	Missense Variant	0.003	0	0	0	0
ESR1	rs188957694	6:152265353	A	Missense Variant	0.0006	0	0	0	0
ESR1	rs2234693	6:152163381	T	Intron Variant	0.395	0.651	0.600	0.578	0.579
ESR1	rs9340799	6:152163335	G	Intron Variant	0.285	0.272	0.194	0.308	0.373
GSTP1	rs1138272	11:67353579	T	Intron Variant	0.004	0.029	0.001	0.071	0.071
GSTP1	rs1695	11:67352689	A	Missense Variant	0.427	0.525	0.821	0.670	0.706
GSTP1	rs199949339	11:67353669	C		0.0012	0	0	0	0
SLC19A1	21:46951785	21:46951785	T	Stop Gained	0.0006	0	0	0	0
SLC19A1	rs1051266	21:46957794	C	Missense Variant	0.307	0.582	0.474	0.549	0.594
SLC19A1	rs202074305	21:46951987	G	Missense Variant	0.0006	0	0	0	0
SLC19A1	rs557998866	21:46935834	C	Missense Variant	0.0018	0	0	0	0
TYMS	rs559346114	18:670817	A	Missense Variant	0.0006	0	0	0	0
TYMS	rs151264360	18:673444-673449	TTAAAG	3 Prime UTR Variant	0.393	0.680	0.306	0.671	0.529
XRCC1	rs138284081	19:44058851	T	Missense Variant	0.002	0.001	0.001	0	0
XRCC1	rs143917286	19:44056954	A	Missense Variant	0.004	0.001	0	0	0
XRCC1	rs1799782	19:44057574	A	Missense Variant	0.092	0.117	0.282	0.052	0.110
XRCC1	rs2307177	19:44047825	G	Missense Variant	0.045	0.030	0.001	0.011	0.017
XRCC1	rs25487	19:44055726	T	Missense Variant	0.118	0.313	0.235	0.366	0.344
XRCC1	rs530200061	19:44057823	T	Missense Variant	0.0006	0	0	0	0
XRCC1	rs532702063	19:44079076	T	Missense Variant	0.0006	0	0	0	0
XRCC1	rs537236859	19:44065100	A	Missense Variant	0.0006	0	0	0	0
XRCC1	rs552435216	19:44079106	A	Missense Variant	0.0006	0	0	0	0
XRCC5	rs1051685	2:216205653	G	3 Prime UTR Variant	0.324	0.108	0.068	0.110	0.170

7.3 Appendix Results Additions

Special Digital Files Access

Large file sets, for the sake of clarity and space saving, are available at the following digital storage location:

https://drive.google.com/drive/folders/1uqum6q9_SULWD51gs4hs0DaoBMsAHPAL?usp=sharing

7.3.1 Non-Coding/Regulatory Region Variants with high scoring predictions

The list of high scoring non-coding and regulatory region variants, along with extra annotations is stored above. The list can be easily browsed using filters above the respective column headers.

7.3.2 African Pairwise statistical comparisons

Due to the size of the dataset, records of the statistical comparisons of African populations compared to each other are stored digitally. These pages are sorted into the ranks of those that are not significantly different from each other, and those that are significant at a level where $p<0.05$

Appendix Table 4: Statistical comparisons of KGP population frequencies compared pairwise against continental African totals

Gene	ID	Location	AMR	EAS	EUR	SAS
ABCB1	7:87195405	7:87195405	1	0.529	0.532	1
ABCB1	rs1045642	7:87138645	$2.51X10^{-51}$	$3.02x10^{-53}$	$3.11x10^{-100}$	$6.13x10^{-125}$
ABCB1	rs1128503	7:87179601	$4.37x10^{-47}$	$9.58x10^{-160}$	$1.10x10^{-62}$	$1.27x10^{-135}$
ABCB1	rs137996914	7:87133567	1	1	1	1
ABCB1	rs142600685	7:87179256	1	1	1	1
ABCB1	rs147487745	7:87150189	1	1	1	1
ABCB1	rs148718120	7:87150095	0.326	0.304	0.304	0.304
ABCB1	rs200754866	7:87165800	1	1	1	1
ABCB1	rs201159898	7:87178825	1	1	1	1
ABCB1	rs2032582	7:87160618	$1.69x10^{-111}$	$5.33x10^{-149}$	$3.98x10^{-155}$	$3.29x10^{-261}$
ABCB1	rs2032583	7:87160561	$3.94x10^{-09}$	$1.16x10^{-24}$	$3.59x10^{-07}$	0.0492
ABCB1	rs2229109	7:87179809	$3.34x10^{-05}$	0.0492	$1.18x10^{-08}$	0.126
ABCB1	rs35730308	7:87138758	0.56	0.293	0.294	0.299
ABCB1	rs535338471	7:87179198	1	1	1	1
ABCB1	rs572038993	7:87145971	1	1	1	1
ABCB1	rs930790323	7:87179284	1	1	1	1
CYP19A1	rs1008805	15:51549599	$1.04x10^{-84}$	$3.89x10^{-17}$	$4.54x10^{-58}$	$5.93x10^{-56}$
CYP19A1	rs142652579	15:51514707	1	1	1	1
CYP19A1	rs143562020	15:51510760	1	1	1	1
CYP19A1	rs4646	15:51502844	$1.20x10^{-15}$	0.965	0.693	$4.35x10^{-07}$
CYP19A1	rs6493497	15:51630835	$4.67x10^{-09}$	$1.57x10^{-08}$	$4.16x10^{-19}$	0.142
CYP19A1	rs749292	15:51558731	$3.99x10^{-05}$	$1.51x10^{-4}$	0.145	$2.89x10^{-09}$
CYP1B1	rs1056836	2:38298203	$3.56x10^{-123}$	$7.48x10^{-300}$	$9.50x10^{-94}$	$1.81x10^{-225}$
CYP1B1	rs1800440	2:38298139	$6.24x10^{-38}$	0.797	$1.54x10^{-73}$	$3.11x10^{-75}$
CYP1B1	rs747324108	2:38298091	1	1	1	1
CYP1B1	rs9282671	2:38302291	1	0.304	0.116	1
CYP2D6	rs1058172	22:42523528	0.0656	0.016	0.0161	0.0311
CYP2D6	rs1065852	22:42526694	$3.78x10^{-05}$	$3.94x10^{-163}$	$3.73x10^{-16}$	$1.31x10^{-08}$
CYP2D6	rs143145507	22:42523891	1	0.529	0.529	0.532
CYP2D6	rs147960066	22:42523592	1	1	1	1
CYP2D6	rs16947	22:42523943	$1.52x10^{-32}$	$5.62x10^{-127}$	$1.26x10^{-36}$	$6.15x10^{-31}$
CYP2D6	rs28371703	22:42525821	$1.43x10^{-20}$	0.00173	$7.66x10^{-56}$	$7.15x10^{-19}$
CYP2D6	rs369177208	22:42522721	0.326	0.304	0.304	0.304
CYP2D6	rs373243894	22:42526672	1	1	1	1
CYP2D6	rs376636053	22:42525158	0.505	1	1	1
CYP2D6	rs3892097	22:42524947	$6.99x10^{-13}$	$6.96x10^{-13}$	$2.16x10^{-32}$	$1.86x10^{-10}$
CYP2D6	rs532668079	22:42522748	1	1	1	1
CYP2D6	rs536109057	22:42526640	1	0.529	0.529	0.532
CYP2D6	rs59421388	22:42523610	$4.32x10^{-28}$	$1.25x10^{-42}$	$1.38x10^{-42}$	$1.97x10^{-41}$

CYP3A4	7:99375691	7:99375691	1	1	1	1
CYP3A4	rs12721629	7:99359800	0.124	0.00902	0.00903	0.00917
CYP3A4	rs2242480	7:99361466	$1.07x10^{-87}$	$6.43x10^{-179}$	$<2.2x10^{-16}$	$1.05x10^{-118}$
CYP3A4	rs2740574	7:99382096	$1.96x10^{-191}$	$<2.2x10^{-16}$	$<2.2x10^{-16}$	$5.35x10^{-312}$
CYP3A4	rs35599367	7:99366316	$1.34x10^{-07}$	0.293	$2.44x10^{-18}$	0.0864
CYP3A4	rs4646440	7:99360870	$2.40x10^{-24}$	$3.46x10^{-24}$	$2.73x10^{-23}$	$2.20x10^{-37}$
CYP3A4	rs530166880	7:99358480	1	1	1	1
CYP3A4	rs540702483	7:99361630	1	1	1	1
CYP3A4	rs57409622	7:99367428	0.326	0.304	0.304	0.304
CYP3A4	rs67784355	7:99359829	1	1	1	1
CYP3A4	rs897005504	7:99377646	0.326	0.304	0.304	0.304
CYP3A5	7:99270249	7:99270249	0.326	0.304	0.304	0.304
CYP3A5	7:99270300	7:99270300	1	1	1	1
CYP3A5	rs111371159	7:99260462	1	1	1	1
CYP3A5	rs142823108	7:99264299	1	0.529	0.529	0.532
CYP3A5	rs145774441	7:99260477	1	0.561	1	1
CYP3A5	rs149664815	7:99247731	0.326	0.304	0.304	0.304
CYP3A5	rs188366390	7:99258153	0.0407	0.00902	0.0372	0.00917
CYP3A5	rs200579169	7:99273810-99273811	0.505	1	$7.17x10^{-05}$	0.0673
CYP3A5	rs41303343	7:99250393-99250394	$4.40x10^{-23}$	$1.31x10^{-35}$	$1.43x10^{-35}$	$8.95x10^{-35}$
CYP3A5	rs776746	7:99270539	$8.09x10^{-169}$	$7.37x10^{-159}$	$<2.2x10^{-16}$	$5.60x10^{-131}$
DPYD	COSM1688098	1:97847978	1	1	1	1
DPYD	rs114096998	1:97544543	$3.72x10^{-06}$	$2.91x10^{-08}$	$2.93x10^{-08}$	$3.43x10^{-08}$
DPYD	rs115232898	1:98165030	$4.44x10^{-4}$	$5.41x10^{-08}$	$5.41x10^{-08}$	$5.95x10^{-08}$
DPYD	rs12022243	1:97862780	0.119	$2.48x10^{-4}$	0.0182	$1.42x10^{-42}$
DPYD	rs142619737	1:97981407	0.326	0.304	0.304	0.304
DPYD	rs144395748	1:98015282	1	0.529	0.529	0.532
DPYD	rs145548112	1:97771751	0.505	1	1	1
DPYD	rs17376848	1:97915624	$6.54x10^{-21}$	$1.04x10^{-37}$	$3.09x10^{-08}$	$1.69x10^{-07}$
DPYD	rs1801158	1:97981421	$1.25x10^{-4}$	1	$1.33x10^{-12}$	0.0295
DPYD	rs1801159	1:97981395	$1.63x10^{-07}$	$4.92x10^{-09}$	0.249	$1.25x10^{-10}$
DPYD	rs1801160	1:97770920	0.0135	0.00737	0.0291	$3.48x10^{-09}$
DPYD	rs1801265	1:98348885	$3.32x10^{-21}$	$4.99x10^{-87}$	$8.07x10^{-30}$	$6.25x10^{-17}$
DPYD	rs201268750	1:97544676	1	1	1	1
DPYD	rs2297595	1:98165091	0.0255	$3.60x10^{-14}$	0.102	0.00701
DPYD	rs573299212	1:98058829	1	0.529	0.529	1
DPYD	rs60511679	1:97770919	1	0.529	0.529	0.532
DPYD	rs61622928	1:98039437	$7.71x10^{-20}$	$1.10x10^{-32}$	$1.18x10^{-32}$	$6.96x10^{-32}$
DPYD	rs72549310	1:98348909	1	1	1	1
DPYD	rs72975710	1:98015291	0.330	0.164	0.164	0.164
ESR1	rs188957694	6:152265353	1	1	1	1
ESR1	rs2234693	6:152163335	$9.74x10^{-30}$	$1.76x10^{-24}$	$1.01x10^{-19}$	$1.39x10^{-19}$

ESR1	rs9340799	6:152163381	$6.15x10^{-01}$	$2.10x10^{-07}$	$1.73x10^{-01}$	$1.77x10^{-06}$
GSTP1	rs1138272	11:67353579	$2.08x10^{-06}$	0.272	$3.00x10^{-23}$	$5.67x10^{-23}$
GSTP1	rs1695	11:67352689	$1.91x10^{-05}$	$6.18x10^{-94}$	$5.01x10^{-34}$	$3.47x10^{-44}$
GSTP1	rs199949339	11:67353669	1	0.529	0.529	0.532
SLC19A1	21:46951785	21:46951785	1	1	1	1
SLC19A1	rs1051266	21:46957794	$5.01x10^{-35}$	$8.03x10^{-18}$	$9.38x10^{-35}$	$3.43x10^{-47}$
SLC19A1	rs202074305	21:46951987	1	1	1	1
SLC19A1	rs557998866	21:46935834	0.56	0.293	0.294	0.299
TYMS	rs559346114	18:670817	1	1	1	1
TYMS	rs151264360	18:673444-673449	$8.78x10^{-4}$	$3.19x10^{-52}$	$9.09x10^{-4}$	$1.03x10^{-4}$
XRCC1	rs138284081	19:44058851	1	1	0.294	0.299
XRCC1	rs143917286	19:44056954	0.681	0.089	0.0891	0.0904
XRCC1	rs1799782	19:44057574	0.0818	$2.16x10^{-36}$	$1.19x10^{-4}$	0.137
XRCC1	rs2307177	19:44047825	0.109	$1.47x10^{-14}$	$3.20x10^{-07}$	$1.46x10^{-4}$
XRCC1	rs25487	19:44055726	$1.70x10^{-27}$	$8.32x10^{-15}$	$1.96x10^{-50}$	$1.61x10^{-42}$
XRCC1	rs530200061	19:44057823	1	1	1	1
XRCC1	rs532702063	19:44079076	1	1	1	1
XRCC1	rs537236859	19:44065100	1	1	1	1
XRCC1	rs552435216	19:44079106	1	1	1	1
XRCC5	rs1051685	2:216205653	$1.838815x10^{-30}$	$1.503934x10^{-59}$	$3.122443x10^{-38}$	$1.204644x10^{-18}$