Characterisation of variation in the *CYP2C19* gene in African populations

Ross Peter Booyse

Student number: 1827129

A research report submitted to the Faculty of Health Sciences, University of the Witwatersrand, in partial fulfilment of the requirements for the Degree of

MSc (Med) Genomic Medicine

6 June 2023

Supervisor: Dr. David Twesigomwe

Co-supervisor: Prof. Scott Hazelhurst

DECLARATION

I, Ross Peter Booyse declare that this Research Report is my own, unaided work. It is being submitted for the degree of MSc (Med) Genomic Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.


6 June 2023

Presentations arising from this research project

1. Booyse R.P., Twesigomwe D., and Hazelhurst S. Characterisation of variation in the *CYP2C19* gene in African populations. Faculty of Health Sciences research day: Oral presentation. 15 September 2022

Abstract

Cytochrome P450 2C19 (*CYP2C19*) is highly polymorphic, and this accounts in part for variability in response to clopidogrel and anti-depressants worldwide. However, *CYP2C19* variation in African populations has been understudied to date. This study aimed to characterise the distribution of *CYP2C19* star alleles (haplotypes) across diverse African populations. We used the StellarPGx pharmacogenomics pipeline to call *CYP2C19* star alleles from 604 secondary genomes representative of six Sub-Saharan African (SSA) populations. We identified 13 previously known *CYP2C19* star alleles in SSA and our analysis revealed 14 predicted novel star alleles across 3% of the participants. CYP2C19 phenotype predictions indicated that 64-73% of SSA participants would require dosage or therapy alterations based on clopidogrel clinical guidelines. Our results emphasize the importance of pharmacogenomics studies for key genes such as *CYP2C19* for precision medicine in the African context. We recommend addition of novel alleles identified in this study to pharmacogenetic testing panels after suitable validation.

Acknowledgements

I would like to thank my supervisor David Twesigomwe, for the never ending support and guidance throughout my research report. Thank you for taking valuable time to guide me through every step and ensuring that I always learn something along the way. The passion you express for this work has rubbed off on me and it has no doubt fired my passion for the field of pharmacogenomics. To my co-supervisor Scott Hazelhurst, thank you for all the help you have provided in getting the work to where it needed to be. Your insights always provide a necessary cornerstone to the science behind the research.

Thank you to the other members at the Sydney Brenner Institute for Molecular Biosciences that have always been there when I needed guidance. Your help and even conversations will always be remembered

Many thanks to Wits University for the Postgraduate Merit Award; this made this year finically possibly and allowed me to meet new scientists learn new skills within the division of Human Genetics.

I would also like to give a special thanks to everyone at the Division of Human Genetics at the NHLS for all the lessons and important knowledge that has helped me grow in my career as a scientist.

As a last thanks, I would like to mention the overwhelming support from my family, friends and the division and my girlfriend. You all have carried me through this year and have been there in my worst times to get me through. This year would not be possible without your support and encouragement.

Table of contents

List of Figures

List of Tables

List of abbreviations

ADRs	Adverse Drug reactions
PGx	Pharmacogenomics
ADME	Absorption, Distribution, Metabolism, and Excretion
CYP	Cytochrome P450
WGS	Whole Genome Sequencing
HTS	High Throughput Sequence
PharmVar	Pharmacogene Variation Consortium
CPIC	Clinical Pharmacogenetics Implementation Consortium
SNPs	Single Nucleotide Polymorphisms
SNVs	Single Nucleotide Variants
Indels	Insertions/Deletions
SSA	Sub-Saharan African
SAS	South Asian
EAS	East Asian
AMR	Admixed American
EUR	European
YRI	Yoruba in Ibadan
ESN	Esan in Nigeria
LWK	Luhya in Webuye, Kenya
GWD	Mandinka in western divisions of The Gambia
MSL	Mende in Sierra Leone
SEB	Southeastern Bantu
NGS	Next-Generation Sequencing
AF	Allele Frequency
IGV	Integrative Genomics Viewer

Research report
in “submissible” paper format

Characterisation of variation in the *CYP2C19* gene in African populations

Ross Peter Booyse[1], David Twesigomwe[1,2] and Scott Hazelhurst[2,3]

[1] Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2] Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[3] School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

Abstract for journal submission

Background: *CYP2C19* pharmacogenetic testing is important clinically to optimise patient response to clopidogrel and anti-depressants. This study aimed to characterise the distribution of *CYP2C19* star alleles (haplotypes) across diverse African populations compared with global populations, with a view towards informing future pharmacogenetic implementations. **Methods:** *CYP2C19* star alleles and diplotypes were called from 604 high coverage genomes from continental African populations using the StellarPGx pipeline. **Results:** From our analysis, *CYP2C19*1* (51%), **2* (17%), and **17* (22%) were the most common star alleles across African populations in this study. We also identified 3% of African participants that had potentially novel *CYP2C19* haplotypes. Over 70% of the SSA participants had either poor, intermediate, rapid, and ultrarapid metabolizer status, and would likely benefit from dosage and/or treatment alterations, especially for clopidogrel **Conclusion:** This study supports the necessity for *CYP2C19* pharmacogenetic testing in African clinical settings and the importance of comprehensive star allele characterisation in the African context.

Introduction

Genetic diversity partly accounts for the variability in drug response within and between populations. This response may range from therapy non-response to severe adverse drug reactions (ADRs) [1,2,3]. Therefore, consideration of patients' genomic profiles when choosing treatment options has the potential to increase medication safety and efficacy in clinical settings globally [4,5]. Implementation of genotype-guided treatment as part of precision medicine, however, requires comprehensive characterisation of variation in genes that encode enzymes involved in drug absorption, distribution, metabolism, and excretion (ADME) [6,7].

Of the ADME genes, cytochrome P450 (CYP) genes are widely studied as they encode enzymes responsible for the metabolism of ~80% of commonly prescribed medications [2,8]. However, there is a paucity of information on CYP variation (and ADME pharmacogenetic variation in general) across African populations owing to their relative underrepresentation compared to European populations in genomics studies to date [9,10]. Our recent ADME pharmacogenetic studies have leveraged the increasing availability of high-depth African genomes [11,12] to comprehensively characterise variation in *POR* [13] and *CYP2D6* [14] across African populations. Given the considerable novel African-ancestry variation we found in these genes, and the extensive genetic diversity in Africa, pharmacogenetic studies are warranted for other key ADME genes.

Based on literature reviews by our group and others [15], we have identified the *CYP2C* gene cluster (*CYP2C8, CYP2C9, CYP2C18* and *CYP2C19*) as the next focus for investigating pharmacogenetic diversity across African populations. This is largely due to the high number of pharmacogenetic clinical annotations, particularly in the context of anti-clotting, anti-inflammatory and anti-depressant drug metabolism and response [16]. Herein, we present a sub-study of this analysis which entailed characterising the *CYP2C19* pharmacogenetic variation across African populations. CYP2C19 is vital in the metabolism of important anti - depressants, and antiplatelet medication such as clopidogrel [17]. Response to these medications ranges from treatment non-response to severe adverse drug reactions and is largely dependent on genetic variation in *CYP2C19* [18]. Based on the increase in the burden of depressive disorders, cardiovascular disease and other non-communicable diseases in Africa [19], *CYP2C19* pharmacogenetic studies are highly relevant in the African context. This is also in keeping with

recommendations from a review by Alessandrini and Pepper [16].

CYP2C19 (94kb-long) consists of nine exons and is located in the *CYP2C* cluster on chromosome 10q23 alongside *CYP2C18*, *CYP2C8* and *CYP2C9* (Figure 1) [18]. The Pharmacogene Variation Consortium (PharmVar) has catalogued 39 star (*) alleles (i.e. haplotypes) for *CYP2C19* to date (https://www.pharmvar.org/gene/CYP2C19 ; last accessed 10-12-22). These star alleles are defined by various combinations of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), and/or structural variations [18,20].

CYP2C19 phenotype categories range from poor metabolizers (PM) to ultra -rapid metabolizers (UM) largely depending on the distribution of *CYP2C19*1* and **38* (normal function), **2* through **4* (non-functional) and **17* (increased function). However, the potential collective functional impact of rare and/or population-specific alleles cannot be overstated. The Clinical Pharmacogenetics Implementation Consortium (CPIC) (https://cpicpgx.org/) maintains five up-to-date drug dosing guidelines involving *CYP2C19*, in particular for clopidogrel and anti-depressants. However, these guidelines are largely based on information from patients of European and Asian ancestry. As such, the transferability of these guidelines to African clinical settings may be challenging since frequencies of clinically actionable alleles tend to differ within and between populations [7, 21]. Given the higher extent of genetic diversity in African populations, we hypothesize that there may be uncharacterised African-ancestry *CYP2C19* variation, especially across understudied African ethnolinguistic groups. Identifying such variation will in turn support the effective implementation of precision medicine strategies in Africa and across the African diaspora thereby alleviating the burden and costs related to adverse drug reactions for medications such as clopidogrel.

Therefore, this study aimed to characterise the distribution of *CYP2C19* star alleles across diverse African populations based on high coverage whole genome sequence (WGS) data analysis. We also leveraged WGS data from public repositories to perform *CYP2C19* star allele distribution comparisons with other global populations. This study adds to the much-needed knowledge on *CYP2C19* variation, particularly in Sub-Saharan African (SSA) populations, which will inform design of genotyping panels, CYP2C19 functional studies, and subsequent clinical pharmacogenetics implementation on the African continent and across the African diaspora.

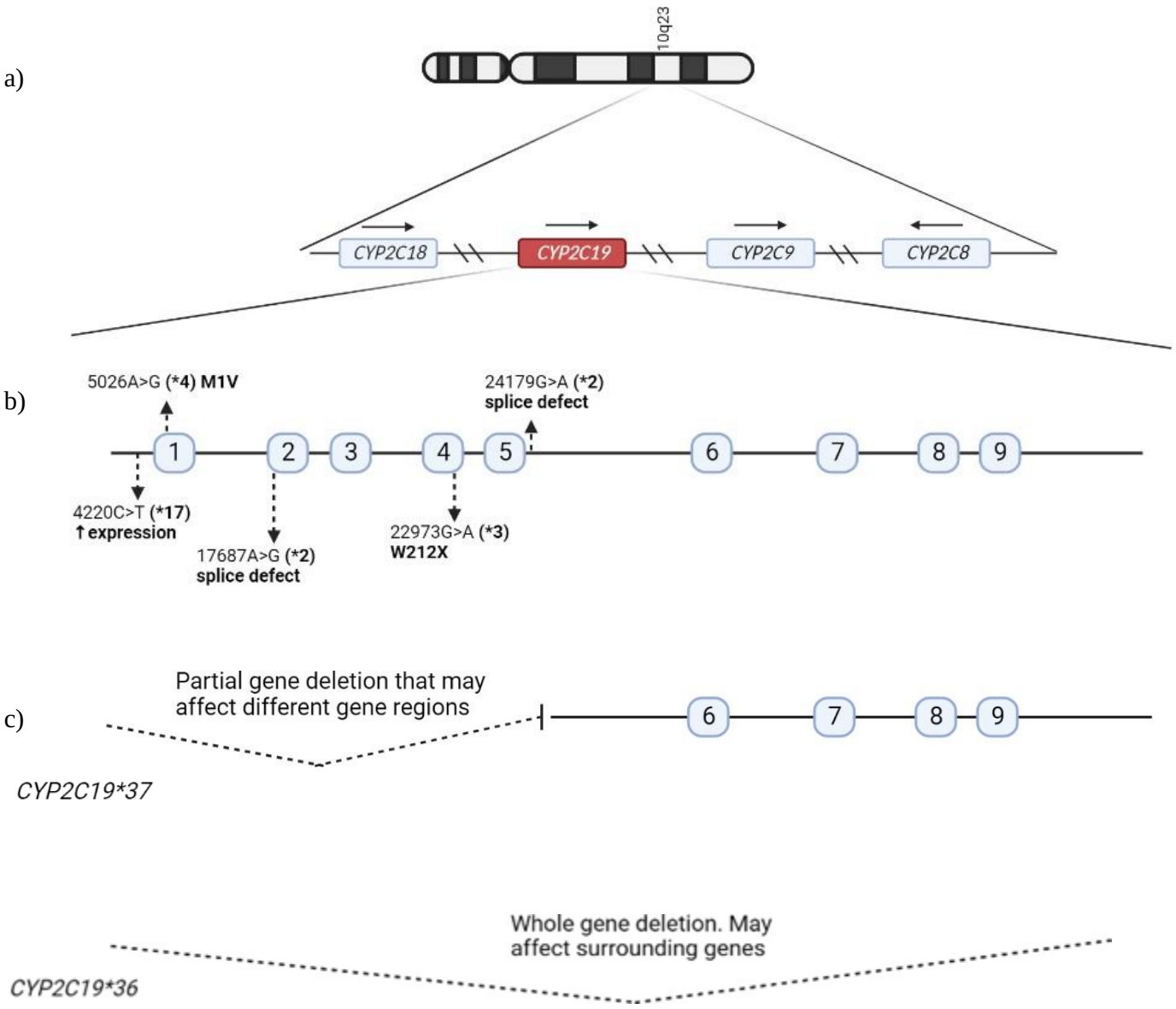

Figure 1: Overview of the *CYP2C* gene locus. (a) Graphical summary of location of genes on the *CYP2C* cluster on chromosome 10. This includes *CYP2C18, CYP2C19, CYP2C9* and *CYP2C8*. Arrows indicate whether the genes are on the forward or reverse DNA strands. (b) Core variants which define *CYP2C19*2* to **4* and **17*. These are previously known common *CYP2C19* star alleles across multiple populations. (c) Examples of *CYP2C19* structural variations that have been described in literature to date. Numbered boxes represent exons. Variant positions are relative to the RefSeq gene NG_008384.3

Materials and Methods

Ethical conduct of research

Ethics approval for this study was granted by the Human Research Ethics Committee (HREC Medical) of the University of the Witwatersrand (Protocol number: M220293). This study involved secondary analysis of existing genomic data only, including public data from the 1000 Genomes Project and WGS data from the Africa Wits-INDEPTH Partnership for Genomic Research (AWI-Gen) study. Permission to use data from the AWI-Gen study was granted by the Principal Investigator: Michele Ramsay. The AWI-Gen study was approved by the Wits HREC Medical under protocol M121029 and M170880.

Population Data

We analysed high-coverage (30x) WGS data in six SSA ethnolinguistic groups represented in the latest 1000 Genomes Project data release [11] and from the AWI-Gen study [22]. The 1000 Genomes Project continental African participants include 108 individuals from the Yoruba in Ibadan (YRI) in Nigeria; 99 individuals from the Esan (ESN) in Nigeria; 99 individuals from the Luhya in Webuye, Kenya (LWK); 85 individuals from the Mende in Sierra Leone (MSL); and 113 individuals from the Mandinka in western divisions of The Gambia (GWD). The AWI-Gen study included South African genomes from 100 southeastern Bantu [SEB]-speaking participants. The total sample size for this study is therefore 604 participants.

WGS datasets (n=2000) from other global populations represented in the 1000 Genomes Project dataset were incorporated into the study for comparison of *CYP2C19* star allele distributions. These include 157 African American/Afro–Caribbean, 503 European (EUR), 504 East Asian (EAS), 489 South Asian (SAS), and 347 admixed American (AMR) participants.

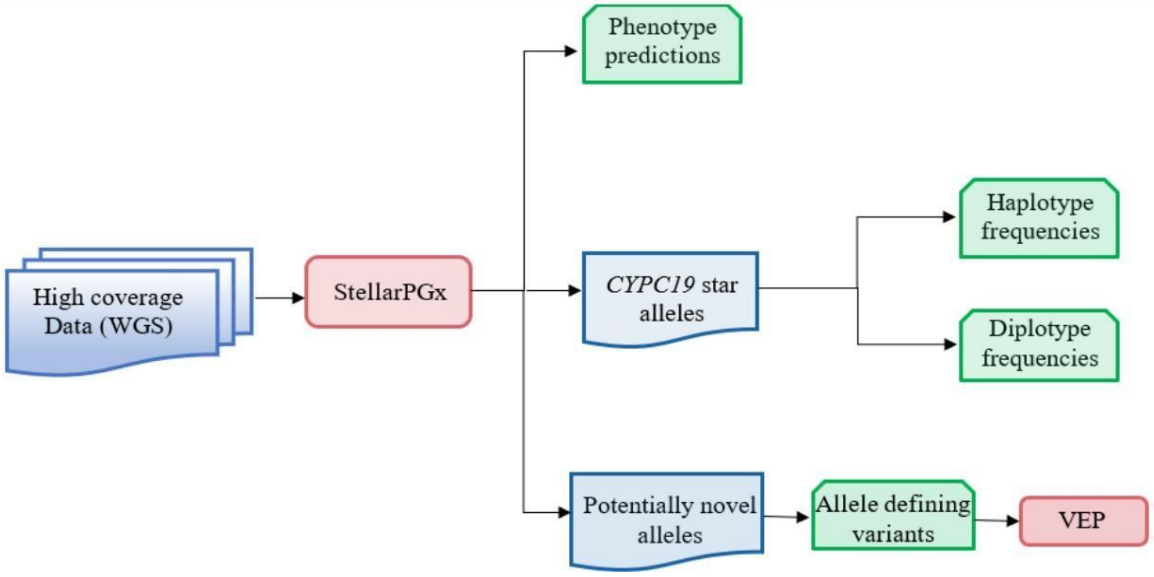

Figure 2: Workflow for bioinformatic analysis of the high-coverage input WGS data represented in the 1000 Genomes and AWI-Gen datasets. VEP, Ensembl Variant Effect Predictor; WGS, Whole Genome Sequence. Blue colour represents data/data points, red colour represents tools and green colour represents key results.

CYP2C19 star allele calling

CYP2C19 star alleles were called using StellarPGx [23] (Figure 2). StellarPGx is a Nextflow-based pipeline that provides a scalable and reproducible approach to determining allelic variation in *CYP* genes according to the star allele nomenclature curated by PharmVar. There was no requirement for preselection of variations to be used in star allele assignment in this study since the input participant data was WGS. StellarPGx applied quality by depth (QD) >10 and 0.25 < allele balance for heterozygous calls (ABHet) < 0.75 among its variant calling filters to improve accuracy in the subsequent *CYP2C19* star allele assignment steps. The current StellarPGx workflow version 1.2.6.1 (https://github.com/SBIMB/StellarPGx) uses high-coverage WGS binary alignment map (BAM) files as inputs (Figure 2). Single nucleotide variations (SNVs) in the gene of interest are identified using GraphTyper [24]. Genome graph based approaches have been shown to be more reliable in variant discovery due to improved read realignment and reduction in reference allele bias. Following variant detection, StellarPGx assigns candidate star alleles (based on core variants in the sample) and then combines this information with copy number changes (if any) detected based on read-depth output from SAMtools [25] to produce final *CYP2C19* diplotypes [23].

Core variants in pharmacogenes like *CYP2C19* typically include missense variants, frameshift variants, stop gain/loss and splice site variants and, occasionally regulatory variants. Importantly, StellarPGx is not affected by population structure as it does combinatorial diplotype assignment rather than statistical phasing [23].

For structural variations, StellarPGx uses read depth distributions and variant allele depth ratios to assess whether (1) a sample contains a full or partial gene deletion (2) which star allele is duplicated in case of copy number > 2. Known *CYP2C19* structural variants assessed by StellarPGx include **36* (full *CYP2C19* deletion) and **37* (partial *CYP2C19* deletion).

For individuals with novel star alleles, StellarPGx includes a flag in the output indicating the presence of a potential novel allele (as none of the diplotypes in the StellarPGx database would be a match) and recommends extra manual variant inspection, and experimental validation according to PharmVar criteria. Novel star alleles could be due to occurrence of novel non-synonymous variants on common haplotype "backbones" (e.g. *CYP2C19*1*) or novel combinations of existing *CYP2C19* core variants defining other haplotypes.

CYP2C19 variant annotation

The Ensembl Variant Effect Predictor (VEP) [26] was used for variant annotation. The VEP plug-ins SIFT [27], Polyphen-2 [28], CADD [29], LRT [30], PROVEAN [31], LOFTEE [32] and VEST4 [33] were used to predict deleteriousness of novel and known allele -defining *CYP2C19* variants. The combination of these tools provides a more accurate prediction for variant impacts in pharmacogenes such as *CYP2C19,* especially with the application of ADME-optimised thresholds [34]. A variant was considered to have a consensus deleterious prediction if at least half of the VEP plugins used to annotate the variant predicted a deleterious impact on CYP2C19.

CYP2C19 phenotype predictions

CYP2C19 metabolizer phenotypes were predicted based on diplotype information from the CPIC guidelines for clopidogrel [1]. Table 1 shows how this phenotype prediction was implemented.

Table 1: Assignment of predicted CYP2C19 phenotypes based on diplotypes

Predicted Phenotype	Genotype	Examples of *CYP2C19* diplotypes
CYP2C19 ultrarapid metabolizer	An individual carrying two increased function alleles	**17/*17*
CYP2C19 rapid metabolizer	An individual carrying one normal function allele and one increased function allele	**1/*17*
CYP2C19 normal metabolizer	An individual carrying two normal function alleles	**1/*1*
CYP2C19 likely intermediate metabolizer	An individual carrying one normal function allele and one decreased function allele or one increased function allele and one decreased function allele or two decreased function alleles	**1/*9, *9/*17, *9/*38*
CYP2C19 intermediate metabolizer	An individual carrying one normal function allele and one no function allele or one increased function allele and one no function allele	**1/*2, *1/*35, *2/*17, *3/*17*
CYP2C19 likely poor metabolizer	An individual carrying one decreased function allele and one no function allele	**2/*9, *2/*10*
CYP2C19 poor metabolizer	An individual carrying two no function alleles	**2/*2, *2/*35, *35/*35*
Indeterminate metabolizer	An individual carrying one or two uncertain function alleles	**1/*39, *2/*12, *17/*39*

Table adapted from Lee et al. (2022).

**1* = normal function

**17* = increased function

*9, **10* = decreased function

**2, *3, *35* =no function

**12, *39* = uncertain function

Statistical analysis

The summary output from StellarPGx was used to compute star allele and diplotype frequencies. The Fishers Exact test was used to determine any significant differences in star allele frequencies between different populations. For correction of multiple testing, we applied the false discovery rate (FDR) p-value adjustment for dependent observations ('BY' method) suggested by Benjamini & Yekutieli [35] for *CYP2C19*1, *2* and **17*. $p < 0.05$ indicated a significant difference. Possible deviations from Hardy–Weinberg equilibrium were investigated using the *Genetics* package (v1.3.8.1.3) in R v4.1.0 (https://www.r-project.org/). $p > 0.05$ indicated that the allele distributions followed Hardy–Weinberg equilibrium expectations.

Results

In the SSA populations we identified 3653 unique variants (~35 variants per kb) in *CYP2C19* (±5kb flanking regions). Of these variants, 48 (1.3%) were found to be coding variants according to the RefSeq transcript NM_000769.4. In the intronic region we identified 3605 (98.6%) variants, with 79 variants being found in the promoter region (NC_000010.11: g.94760741-94762704). The breakdown for all these variants (by type) is shown in Figure 3.

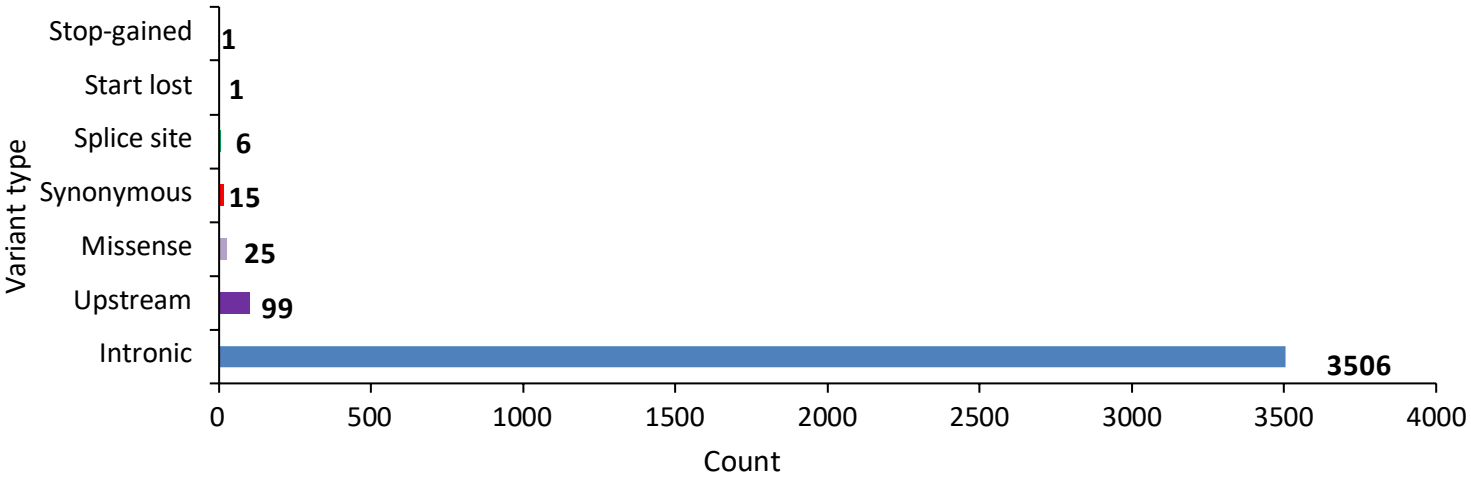

Figure 3. Breakdown of *CYP2C19* variant annotations obtained from the Ensembl Variant Effect Predictor according to the NM_000769.4 transcript in Sub-Saharan African populations in this study.

CYP2C19 star allele frequency distributions

We identified 13 previously known *CYP2C19* star alleles across all SSA populations in this study. *CYP2C19*1, *2*, and **17* were the most commonly observed star alleles (Table 2 and Table 3). The *CYP2C19*2* (loss-of-function allele) frequency was highest in the ESN (20%) and lowest in the GWD (13%) participants, while the *CYP2C19*17* (increased function allele) frequency was highest in the ESN (26%) and lowest in the SEB (17%) participants. However, there were no significant differences in star allele frequencies between any of the SSA ethnolinguistic groups in this study as determined by the Fisher's exact test. Apart from the commonly observed star alleles, other star alleles observed in SSA had low frequencies and some were present as singletons (Table 2 and Table 3). For all star alleles identified there were no significant deviations from Hardy Weinberg Equilibrium (HWE), indicating that there was no underlying selection.

In our global comparison, *CYP2C19*1* occurred at a significantly higher frequency in the Admixed American (AF=73%, p-value=3.9×10^{-10}) populations and a significantly lower frequency in the South Asian (AF=39%, p-value=8.3×10^{-5}) populations when compared to SSA (Table 2). Additionally, *CYP2C19*2* occurred at significantly lower frequency in the African American/Afro-Caribbean (AF=15%, p-value=0.006) populations and occurred at higher frequencies in the South Asian (AF=33%, p-value = 3.5×10^{-5}) and East Asian (AF=31%; p-value = 7×10^{-7}) populations when compared to SSA (Table 2). Conversely, *CYP2C19*17* was present at a significantly higher frequency in SSA (AF=22%) when compared to admixed American (AF=11%, p-value=0.0004), South Asian (AF=15%, p-value=0.001) and East Asian (AF=2%, p value=4×10^{-15}) populations (Table 2). *CYP2C19*12* and **30* were observed in the SSA populations but were not called in any of the other global populations, while *CYP2C19*4, *8, *11, *22, *24, *34* and **37* were not observed in SSA but were present in various other global populations (Table 2).

Table 2: *CYP2C19* star allele distributions across global populations

				Allele frequencies (%)					
Star allele	**CPIC function**	**Defining variant(s)**	**Consequence**	**SSA**	**African American/ Afro-Caribbean**	**AMR**	**SAS**	**EAS**	**EUR**
**1*	Normal function	rs3758581	Missense (I331V)	51.4	52.7	73.3	38.5	57.2	55.8
**2*	No function	rs12769205, rs4244285	Splice defect Splice defect	17.3	14.7	10.2	32.7	31	14.4
**3*	No function	rs4986893	Stop-gain (W212X)	<0.1	0.3	0	1.3	5.6	0
**4*	No function	rs28399504	Start lost (M1A)	0	0	0.3	0	<0.1	0.1
**8*	No function	rs41291556	Missense (W120R)	0	0.3	0	0.1	0	0.3
**9*	Decreased function	rs17884712	Missense (R144H)	1.4	0.6	0.1	0	0	0
**10*	Decreased function	rs6413438	Missense (P227L)	<0.1	0	0.1	0	0	0
**11*	Normal function	rs58973490	Missense (R150H)	0	0.3	0	0	0	0.2
**12*	Uncertain function	rs55640102	Stop-lost (X491C)	<0.1	0	0	0	0	0
**13*	Normal function	rs17879685	Missense (R410C)	1.8	2.2	0.3	0	0	0
**15*	Normal function	rs17882687	Missense (I19L)	1.4	0.6	0	0	0	0
**17*	Increased function	rs12248560	Regulatory (Expression)	22.2	24.3	11.5	14.0	1.5	22.1
**22*	No function	rs140278421	Missense (R186P)	0	0.3	0.1	0	0	0
**24*	No function	rs118203757	Missense (R335Q)	0	0.3	0	0	0	0
**30*	Uncertain function	rs145328984	Missense (R73C)	<0.1	0	0	0	0	0
**34*	Uncertain function	rs367543002, rs367543003	Missense (P3S) Missense (F4L)	0	0	0	1.2	0	0
**35*	No function	rs12769205	Splice defect	3	1.6	0.1	0	0	0
**37*	No function	-	Partial gene deletion	0	0	0	0	0	0.1
**38*	Normal function	Reference allele	-	0.4	1.3	3.8	11.8	4.5	0.7
**39*	Uncertain function	rs17882687, rs17885179	Missense (I19L) Missense (E122A)	0.9	0.3	0	0	0	0

SSA: Sub-Saharan Africa, AMR: admixed American, SAS: South Asian, EAS: East Asian, EUR: European, CPIC: Clinical Pharmacogenomics Implementation consortium.

Table 3: *CYP2C19* haplotype frequencies across six continental African populations in the study

		Allele frequencies (%)					
Haplotype	**CPIC function**	**SEB**	**YRI**	**MSL**	**LWK**	**GWD**	**ESN**
**1*	Normal function	52	49	53	55	52	48
**2*	No function	19	17	17	18	13	20
**3*	No function	0	0	0	1	0	0
**9*	Decreased function	3	0	2	1	1	1
**10*	Decreased function	0	0	1	0	0	0
**12*	Uncertain function	1	0	0	0	0	0
**13*	Normal function	1	2	1	4	3	0
**15*	Normal function	5	0	0	1	3	1
**17*	Increased function	17	24	24	18	24	26
**30*	Uncertain function	0	0	0	0	0	1
**35*	No function	3	6	2	2	3	2
**38*	Normal function	1	0	1	0	0	1
**39*	Uncertain function	0	2	0	0	1	2

SEB: Southeastern Bantu; YRI: Yoruba in Ibadan; MSL: Mende in Sierra Leone; LWK: Luhya in Webuye, Kenya; GWD: Mandinka in western divisions of The Gambia; ESN: Esan in Nigeria; CPIC: Clinical Pharmacogenomics Implementation consortium.

CYP2C19 phenotype distributions

Approximately 5% of the SSA individuals were predicted to be poor CYP2C19 metabolizers (PM). The PM phenotype was less common (~1%) across EUR, AMR and African American/Afro-Caribbean participants. The PM phenotype was observed at a higher frequency in the East and South Asian populations, however this was not statistically different (p-value = 0.08) from the SSA participants (Figure 4). The likely poor metabolizers (LPM) were only observed in the SSA participants.

For the rapid metabolizer (RM) phenotype group there were also no significantly different frequencies across SSA. The frequency across SSA was relatively high (30%) which is mainly accounted for by the **1/*17* diplotype (Supplementary Table 2). Comparing to other global populations, the RM phenotype was observed at a higher frequency in the African American/Afro-Caribbean (39%) and EUR (35%) populations. The frequency of RMs was, however, significantly lower than that observed in the AMR (16%, p -value=0.005), SAS (14%, p-value=0.0003) and EAS (1%, p-value=$3x10^{-7}$) populations when compared to SSA (24%) (Figure 4).

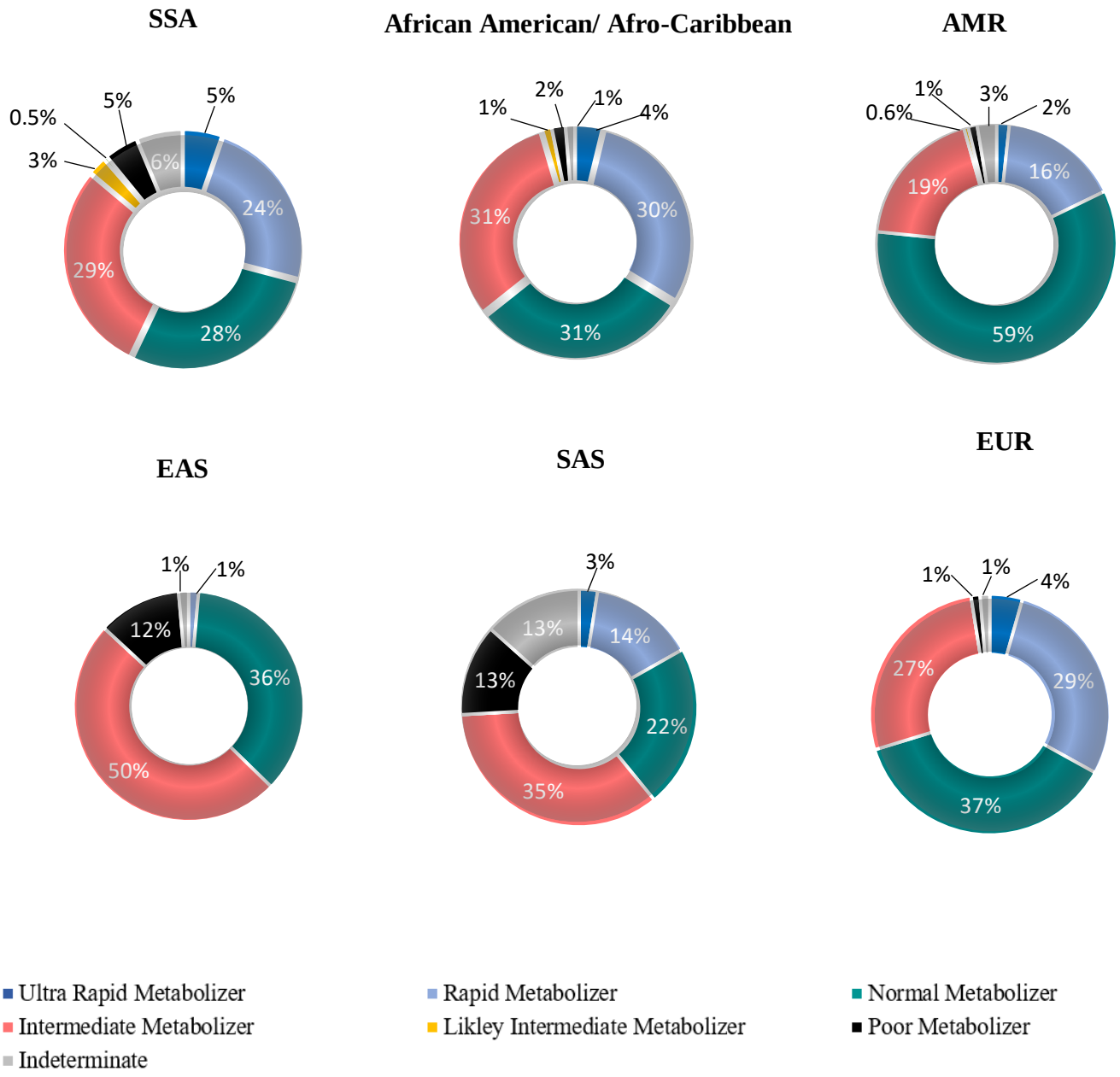

Figure 4: Distribution of predicted phenotypes for global populations in this study. Metabolizer status was determined using the Clinical Pharmacogenetics Implementation Consortium's (CPIC) *CYP2C19* genotype-phenotype translation guidelines. PM; Poor Metabolizer, SSA; Sub-Saharan African participants, AMR; admixed American participants, SAS; South Asian participants, EAS; East Asian participants, EUR; European participants. Phenotype differences in SSA compared to EAS and SAS should be noted, particularly with regards to the rapid, intermediate and poor metabolizers.

We observed the highest frequency of predicted ultrarapid metabolizers (UM) across the SSA populations (5%) when compared with other global populations (UM frequency = 0 -4%) in this study (Figure 4). The frequency of UMs across SSA ranged from 3 -7%. None of the differences in UM phenotype distribution were statistically significant.

In SSA we observed over 6% of participants with indeterminate *CYP2C19* phenotypes. This was lower than the percentage observed in SAS populations (13%), but higher than that in EUR, AMR African American/Afro-Caribbean and EAS populations (Figure 4).

Novel *CYP2C19* alleles inferred in this study

The *CYP2C19* variant positions in this section are based on NG_008384.3, counting from the sequence start, and the protein positions are based on NP_000760.1. The SSA populations had 27 individuals (~4%) who were predicted to have novel haplotypes. This was greater than the number of individuals with predicted novel haplotypes in AMR (~3%), EUR (~1%), EAS (~0.6%) and African American/Afro-Caribbean (~0.6%) participants. However, the SAS populations had a considerably higher proportion of individuals with predicted novel haplotypes (~11%) when compared to SSA (Supplementary Table 3).

We inferred six potential novel *CYP2C19* star alleles in SSA—at least one in each subpopulation, except for YRI (Table 4). Nine participants (~11%) among the MSL had predicted novel *CYP2C19* star alleles. The variants defining the predicted novel haplotypes in this study are already in the Single Nucleotide Polymorphism database (dbSNP, [36]). However, the observed haplotypes in this context have not yet been used to define a star allele by PharmVar and are therefore considered potentially novel—pending final novel star allele or sub allele assignment by the PharmVar *CYP2C19* expert panel.

From the variants defining these novel haplotypes, only rs146991374 (defining haplotype 1 and 4) and rs150152656 (defining haplotype 2) had consensus deleterious VEP plugin predictions (Table 4). Notably, the novel haplotype 5 (defined by rs112197069) was only present in the SEB populations.

Furthermore, haplotypes 1 (defined by rs146991374) and 6 (defined by rs145119820); which were not present in the SEB participants, were present in more than one individual. Haplotype 6 was found in ten different SSA individuals (Table 4). In addition, there were 7 participants for whom StellarPGx predicted presence of potential novel star alleles, however, we could not determine the backbone allele(s) for the defining nonsynonymous variants (Table 4). All of these unresolved diplotypes occurred in singletons. Among these, rs183701923 and rs192154563 in unresolved diplotypes 1 and 2, respectively, had a consensus deleterious prediction. Participant IDs relating to the resolved predicted novel haplotypes and unresolved pr edicted novel diplotypes are provided in Supplementary Table 3. VEP plugin scores for all the *CYP2C19* allele-defining variants called in this study are provided in Supplementary Table 4.

Table 4: Inferred novel CYP2C19 haplotypes and unresolved diplotypes in African populations in this study

Resolved haplotype						
#	**Background allele**	**Additional core variant(s)**	**Consequence(s)**	**Consensus VEP-based prediction**	**Count**	**Country/dataset**
1	**1*	rs146991374~95056~A>T	Missense (K432I)	Deleterious	4	MSL/1000G
2	**1*	rs150152656~ 17767~C>T	Missense (T130M)	Deleterious	1	ESN/1000G
3	**1*	rs145328984~17426~C>T	Missense (R73C)	Deleterious	1	ESN/1000G
4	**1*	rs201306972~ 17427~G>A	Missense (R73H)	Benign	1	MSL/1000G
5	**39*	rs112197069~ 17538~T>A	Missense (F110Y)	Benign	3	SEB/AWI-Gen
6	**2*	rs145119820~ 17715~G>A	Missense (V113I)	Benign	10	MSL,LWK,GWD/1000G
Unresolved diplotypes						
#	**Background allele**	**Additional core variant(s)**	**Consequence(s)**	**Consensus VEP-based prediction**	**Count**	**Country/dataset**
1	[**1/*2*]	rs183701923~ 22893~C>T	Missense (R186C)	Deleterious	1	LWK/1000G
2	[**17/*35*]	rs192154563~ 95085~C>T	Missense (R442C)	Deleterious	1	LWK/1000G
3	[**2/*2*]	rs201306972~ 17427~G>A+ rs537050749~ 24271~A>G	Missense (R73H) Missense (N258S)	Benign	1	ESN/1000G
4	[**1/*17*]	rs375760063~ 17856~A>G	Missense (K160E)	Benign	1	MSL/1000G
5	[**1/*17*]	rs545642100~ 17526~C>G	Missense (A106G)	Benign	1	ESN/1000G
6	[**2/*17*]	rs527499296~ 92332~A>C	Missense (E415D)	Benign	1	LWK/1000G
7	[**1/*3*]	rs144056033~ 92303~A>G	Missense (M406V)	Benign	1	LWK/1000G

Note that *CYP2C19*1* is defined by rs3758581 (I331V). Variant positions are according to the *CYP2C19* RefSeq gene (NG_008384.3).

ID: Identifier, SEB: Southeastern Bantu, MSL: Mende in Sierra Leone, LWK: Luhya in Webuye Kenya, GWD: Mandinka in western divisions of The Gambia, ESN: Esan in Nigeria, 1000G: 1000 Genomes Project, AWI-Gen: Africa Wits-INDEPTH Partnership for Genomic Research.

Discussion

Assessing variation in *CYP2C19* in diverse populations is important as functional variants in this gene might contribute to differences in the efficacy and safety of various important medications such as clopidogrel and anti-depressants. Most of the current literature is focused on the distribution of important star alleles in European and Asian populations while far less is known about African populations [3] despite their notable high genetic diversity. Therefore, although precision medicine strategies such as genotype-guided therapy offer much promise in curbing adverse drug reactions and promoting treatment efficacy, there is a risk of poor transferability of Eurocentric guidelines to African populations which may exacerbate existing health disparities [37]. This study provides important information on the distribution of *CYP2C19* allelic variation in six diverse SSA populations represented in the 1000 Genomes Project and AWI-Gen datasets. Our findings can inform more comprehensive *CYP2C19* pharmacogenetic testing in clinical and research settings. This is pertinent in the African context where there is high genetic heterogeneity and inter-individual response to a number of the relevant aforementioned medications may vary due to *CYP2C19* variation.

*CYP2C19*2* and **17* are among the actionable star alleles in clinical settings worldwide. *CYP2C19*2* (loss-of-function) is associated with a significant decrease in platelet responsiveness to clopidogrel [1] while **17* (increased function) homozygotes are known to be at a greater risk of clopidogrel-related bleeding events [38]. In SSA populations in this study, these two key star alleles occurred at relatively high frequencies similar to those reported in previous studies [38,39,40], see also PharmGKB [41]. *CYP2C19*2* and **17* frequencies did not differ significantly between the African ethnolinguistic groups in this study despite the large genetic diversity in these populations. However, the estimated frequencies, especially for rarer alleles, should be interpreted with caution given the relatively small sample sizes. Across other global populations *CYP2C19*17* was observed at a significantly lower frequency (East Asian populations in particular) when compared to SSA, whereas *CYP2C19*2* was more frequent than that in SSA. This exemplifies the limitation of using other global populations as a proxy for SSA populations when implementing genotype-guided therapy.

Across the SSA populations in this study, 64 -73% of participants had phenotypes (UM, RM, IM, LIM, PM) that would require alterations in clopidogrel and/or anti-depressant dosage or

medication according to current CPIC guidelines [1]. A similar proportion (~60%) of these phenotypes combined has been reported in South African Xhosa, and Cape Mixed Ancestry participants [39]. This exemplifies the need for *CYP2C19* pharmacogenetic testing to minimize adverse drug reactions and promote drug safety across Africa. In our study, phenotype predictions (especially PM, IM and UM) for SSA were significantly different when compared in particular to the SAS and EAS populations. This is in part accounted for by differences in *CYP2C19* allele distributions and emphasizes limitations in transferability of actionable star alleles and genotyping platforms based on non-African populations.

We inferred novel *CYP2C19* haplotypes which may be used guide new star number assignments by PharmVar. In SSA, 3% of participants carried such haplotypes. Four of the eight computationally-resolved novel star alleles were found in more than one individual, which provides more evidence for their validation. In general, potential novel star alleles–and alleles with unknown function–contributed to the non-negligible proportion (6%) of SSA participants with indeterminate CYP2C19 phenotypes. Failure to account for this African-ancestry variation may impact the effective implementation of precision medicine across the continent as African populations may have unique *CYP2C* genetic markers relating to drug response [42]. Furthermore, the calling of novel alleles in this study was facilitated by the use of high coverage WGS data which highlights information that might be missed when using targeted genotyping methods in Africa at present.

This study had some limitations. Firstly, the WGS data used were generated using short read technology. This presents challenges in investigating complex structural variations involving *CYP2C19* and other highly similar *CYP2C* cluster genes (e.g. *CYP2C18*). For example, StellarPGx initially called HG00268 as having *CYP2C19*36* (full gene deletion), however upon visual inspection on the Integrative Genomics Viewer [43] we revised this to *CYP2C19*37* (partial gene deletion) based on the read coverage distribution. Short read data limitations can also account for unresolved singleton novel diplotypes in the study. Secondly, *in silico* variant annotations may not provide accurate variant effect predictions and could not be used to comprehensively assess the function of novel haplotypes. However, molecular functional assays were not within the parameters of this study. Nonetheless, we expect functional effect predictions for stop gain, and splice defect variants to be consistent with experimental assay

validations. Lastly, our phenotype predictions were solely based on the CPIC guidelines outlined for CYP2C19's interaction with clopidogrel. CPIC genotype-phenotype translation for other drugs may have provided different results. Given the importance of clopidogrel in Africa and globally, it provides a justifiable basis for CYP2C19 phenotype prediction comparisons. However, patient response to clopidogrel and other medications could be impacted by other factors such as age, lifestyle, environment, and phenocoversion – whereby genotypic NM patients may present as phenotypic PMs in presence of CYP2C19 inhibitors. We also observed a significant proportion of individuals in SSA with indeterminate CYP2C19 phenotypes (6%). This highlights the shortcomings for the utility of the current CPIC guidelines for genotype–phenotype therapy options in African populations and therefore the need for thorough star allele characterisation and phenotypic research across Africa.

Conclusion

This study contributes to the efforts to characterise the distribution of actionable *CYP2C19* star alleles and associated phenotypes in SSA, and to establish a more comprehensive catalog of potential pharmacogenomically-relevant star alleles from African and other global populations. While no significant *CYP2C19* star allele frequency differences were observed across SSA in this study, our results emphasize key differences across global populations. The relatively large number of predicted novel *CYP2C19* star alleles found in SSA emphasizes the need for thorough pharmacogene star allele characterisation across Africa, especially in understudied ethnolinguistic groups. This could help guide the development of more comprehensive genotyping panels and phenotype prediction algorithms across Africa and for the African diaspora. In turn, such interventions will be critical in improving the efficacy of medications such as clopidogrel and anti-depressants, as well as reducing the risk of adverse drug reactions in patients of African-ancestry.

Future research

Follow-up studies to this work will entail validation (e.g. using long read resequencing) of novel star alleles inferred in this study, and submission to PharmVar for cataloguing, as this is recommended before including them in clinical implementation guidelines. Future studies aimed at determining the CYP2C19 functional impact of variants defining these novel alleles as well as those in unknown function star alleles are also warranted. Furthermore, it is necessary to evaluate the clinical utility and cost-effectiveness of CYP2C19 genotype-guided therapy in the African context given the extent of pharmacogenetic variation described in this study. Considering that this is a pilot study, the use of more diverse population groups in genetic characterisation studies could allow for more informative results, and ultimately better the potential for clinical application improvements.

Data availability statement

The 1000 Genomes Project dataset is publicly available (www.internationalgenome.org/). The 100 high-coverage genomes from the Africa Wits-INDEPTH Partnership for Genomic Research study (EGAD00001006418) are available from the European Genome–Phenome Archive (https://ega-archive.org/).

Code availability statement

The StellarPGx code used in this study can be found at https://github.com/SBIMB/StellarPGx. All custom scripts used in downstream analysis are available at https://github.com/ross-booyse/CYP2C19_PGx.

Author contributions

RP Booyse, D Twesigomwe and S Hazelhurst designed the study. RP Booyse, D Twesigomwe and S Hazelhurst contributed to acquisition and interpretation of the data. RP Booyse analyzed the data. RP Booyse drafted the manuscript. RP Booyse, D Twesigomwe and S Hazelhurst revised the manuscript for important intellectual content. D Twesigomwe and S Hazelhurst supervised the research. All authors approved the final version of the manuscript.

References

1. Lee C R, Luzum J A, Sangkuhl K *et al*. Clinical Pharmacogenetics Implementation Consortium Guideline for CYP2C19 Genotype and Clopidogrel Therapy: 2022 Update. *Clin. Pharm. and Ther.*. *112*(5):959-967 (2022).
2. Zhao M, Ma J, Li M *et al*. Cytochrome P450 Enzymes and Drug Metabolism in Humans. *Int. J. Mol. Sci*.. *22*(23), 12808 (2021).
3. Zhou Y, Ingelman-Sundberg M, Lauschke V M. Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. *Clinical Pharmacol. and Ther. 102*(4), 688–700 (2017).
4. Carr D F, Pirmohamed M. Precision medicine in drug safety. *Curr. Opin. Toxicol.*. *23–24*, 87–97(2020).
5. Desta, Z., Gammal, R. S., Gong, L *et al*. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2B6 and Efavirenz-Containing Antiretroviral Therapy. *Clin Pharmacol Ther*. 106(4), 726-733 (2019).
6. Arbitrio M, Di Martino M T, Scionti F, Barbieri V, Pensabene L, Tagliaferri P. Pharmacogenomic Profiling of ADME Gene Variants: Current Challenges and Validation Perspectives. *High-Throughput 2018. 7*(4), 40 (2018).
7. da Rocha J E B, Othman H, Botha G *et al* The Extent and Impact of Variation in ADME Genes in Sub-Saharan African Populations. *Front. in Pharmacol. 12*, 366 (2021).
8. Zanger U M, Schwab M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol and Ther. 138*(1),103–141 (2013).
9. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin A R, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat. Med. 28*(2), 243–250 (2022).
10. Luczak, T, Stenehjem D, Brown J. Applying an equity lens to pharmacogenetic research and translation to under-represented populations. *Clin. and Transl. Sci. 14*(6), 2117–2123 (2021).
11. Byrska-Bishop M, Evani US, Zhao X et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 185(18), 3246–3440 (2022).

12. Choudhury A, Aron S, Botigué L R *et al.* High-depth African genomes inform human migration and health. *Nature*, *586*(7831), 741–748(2020).
13. Booyse R P, Twesigomwe D, Hazelhurst S. Characterization of POR haplotype distribution in African populations and comparison with other global populations. *Pharmacogenomics*. *23*(14), 771–782 (2022).
14. Twesigomwe D, Drögemöller B I, Wright G E B *et al*. Characterisation of CYP2D6 pharmacogenetic variation in sub-Saharan African populations . *Clin. Pharmacol. & Ther*. (2022).
15. Rajman I, Knapp L, Morgan T, Masimirembwa C. African Genetic Diversity: Implications for Cytochrome P450-mediated Drug Metabolism and Drug Development. *EBioMed*. (2017).
16. Whirl-Carrillo M, Huddart R, Gong L *et al*. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin. Pharmacol*. *3*(110), 563-572 (2021).
17. Alessandrini M, Pepper M S. Priority pharmacogenetics for the African continent: focus on CYP450. *Pharmacogenomics*. *15*(3), 385–400 (2014).
18. Botton M R, Whirl-Carrillo M, Del Tredici A L *et al*. PharmVar GeneFocus: CYP2C19. *Clin. Pharm. & Ther*. *109*(2), 352–366 (2021).
19. Abbafati C, Abbas K M, Abbasi-Kangevari M *et al*. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. *396*(10258), 1204–1222 (2020).
20. Botton M R, Lu X, Zhao G *et al*. Structural variation at the CYP2C locus:Characterization of deletion and duplication alleles. *Hum Mutat*. *40*(11), e37 (2019).
21. Pereira, L, Mutesa L, Tindana P, Ramsay M. African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics*. *22*(5), 284–306 (2021).
22. Ramsay M, Crowther N, Tambo E *et al*. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob. Health Epidemiol. Genom. 1*, 1–13 (2016).

23. Twesigomwe D, Drögemöller BI, Wright GEB ¨ et al. StellarPGx: a Nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clin. Pharmacol. Ther*. 110(3), 741–749 (2021).
24. Eggertsson H P, Jonsson H, Kristmundsdottir S *et al*. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat Genet.* 49, 1654–1660 (2017).
25. Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics 25*(16):2078 (2009).
26. McLaren W, Gil L, Hunt S E *et al*. The Ensembl Variant Effect Predictor. *Genome Biol. 17*(1), 122 (2016).
27. Kumar P, Henikoff S, Ng P C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols. 4*(7), 1073–1082 (2009).
28. Adzhubei I, Jordan, D M, Sunyaev, S R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. 0 7*(SUPPL.76) (2013).
29. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res. 47*(1), 886–894 (2019).
30. Fujita, A, Kojima K, Patriota A G, Sato J R, Severino P, Miyano S. A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets. *Bioinformatics. 26*(18), 2349–2351 (2010).
31. Choi Y, Chan A P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics. 31*(16), 2745–2747(2015).**S**
32. Karczewski K J, Francioli L C, Tiao G *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature. 581*(7809), 434–443 (2020).
33. Carter H, Douville C, Stenson P D, Cooper D N, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genom. 14* (Suppl 3), S3 (2013).
34. Zhou Y, Mkrtchian S, Kumondai M, Hiratsuka M, Lauschke VM. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics* J. 19(2), 115–126 (2019).
35. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Inst. Math. Stat. Collect. 29*(4), 1165-1188 (2001).

36. Smigielski E M, Sirotkin K, Ward M, Sherry S T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res*. *28*(1), 352 (2000).
37. Hussein AA, Hamad R, Newport MJ, Ibrahim ME. Individualized Medicine in Africa: Bringing the Practice Into the Realms of Population Heterogeneity. *Front Genet. 13*(1), 873 (2022).
38. Biswas M. Global Distribution of CYP2C19 Risk Phenotypes Affecting Safety and Effectiveness of Medications. *Pharmacogenomics.* 21:2 21(2): 190–99 (2020).
39. Matimba, A, Del-Favero J, Van Broeckhoven C, Masimirembwa C. Novel Variants of Major Drug-Metabolising Enzyme Genes in Diverse African Populations and Their Predicted Functional Effects. *Hum. Genom.*3(2): 169–90 (2009).
40. Drögemöller BI, Wright GE, Niehaus DJ *et al*. Characterization of the Genetic Profile of CYP2C19 in Two South African Populations. *Pharmacogenomics* 11(8): 1095–1103 (2010).
41. Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. Methods *Mol Biol*. 1015: 311-20 (2013)
42. Ndadza A, Muyambo S, Mntla P et al. Profiling of Warfarin Pharmacokinetics-associated Genetic Variants: Black Africans Portray Unique Genetic Markers Important for an African Specific Warfarin Pharmacogenetics-dosing Algorithm. *J. Thromb. Haemost*. 19(12): 2957 (2021)
43. Robinson J T, Thorvaldsdóttir H, Winckler W et al. Integrative genomics viewer. *Nat. Biotech*. 29(1), 24–26 (2011).

Supplementary Table 1: Diplotype distributions for the global populations included in the study for comparison.

	Diplotype frequencies (%)					
Diplotype	SSA	AMR	SAS	EAS	EUR	African American/ Afro-Caribbean
**1/*1*	26	54	14	32	29	25
**1/*2*	18	15	22	39	18	16
**1/*3*			0.4	6		
**1/*4*		1		0.2	0.2	
**1/*8*			0.2		0.4	1
**1/*9*	2					1
**1/*10*	0.2	0.3				
**1/*13*	2					3
**1/*15*	2					1
**1/*17*	23	15	12	1	26	28
**1/*24*						1
**1/*34*			1			
**1/*35*	4					3
**1/*37*					0.2	
**1/*38*		5	8	4	8	1
**1/*39*	1					1
**2/*2*	4	1	11	8	1	1
**2/*3*			1	4		
**2/*9*	0.3					
**2/*10*	0.2					
**2/*11*					0.4	
**2/*13*	0.3					1
**2/*15*	1					1
**2/*17*	7	3	5	1	6	10
**2/*34*			1			
**2/*35*	1	0.3				
**2/*38*		0.3	7	3	2	
**2/*39*	0.2					
**3/*3*				0.2		
**3/*17*	0.2		0.4	1		
**3/*22*						1
**3/*38*			0.4	1		
**8/*17*					0.2	
**9/*13*		0.3				
**9/*17*	1					1
**9/*38*	0.2					
**11/*17*						1
**12/*15*	0.2					
**13/*17*	1	0.3				1
**13/*35*	0.2					
**15/*17*	1					
**17/*17*	6	2	2	4		4
**17/*22*		0.3				
**17/*34*			1			
**17/*35*	1					1
**17/*38*	0.2	1	2	0.2	3	1
**17/*39*	0.2					
**35/*35*	0.2					
**38/*38*			1	0.2	0.2	

SSA: Sub-Saharan Africa, AMR: admixed American, SAS: South Asian, EAS: East Asian, EUR: European.

Supplementary Table 2: *CYP2C19* diplotype frequencies with corresponding metabolizer status across the African populations in this study.

Diplotype	SEB %	YRI %	MSL %	LWK %	GWD %	ESN %	Metabolizer Status*
**17/*17*	3	7	7	3	5	7	Ultra-Rapid
**1/*17*	18	26	29	19	24	23	Rapid
**13/*17*	1	1	0	2	2	0	Rapid
**15/*17*	2	0	0	0	2	0	Rapid
**17/*35*	0	1	1	0	2	1	Rapid
**17/*38*	1	0	0	0	0	0	Rapid
**1/*1*	25	21	25	30	29	25	Normal
**1/*13*	1	2	1	6	3	0	Normal
**1/*15*	5	0	0	0	3	1	Normal
**12/*15*	1	0	0	0	0	0	Normal
**1/*9*	2	0	3	1	2	2	Likely Intermediate
**1/*10*	0	0	0	0	1	0	Likely Intermediate
**9/*17*	1	0	1	1	1	0	Likely Intermediate
**9/*38*	1	0	0	0	0	0	Likely Intermediate
**1/*2*	23	17	18	21	12	17	Intermediate
**1/*35*	5	8	1	3	2	1	Intermediate
**2/*13*	0	1	0	1	0	0	Intermediate
**2/*15*	1	0	0	1	1	0	Intermediate
**2/*17*	5	6	7	4	6	14	Intermediate
**3/*17*	0	0	0	1	0	0	Intermediate
**13/*35*	0	0	0	0	1	0	Intermediate
**2/*9*	1	1	0	0	0	0	Likely Poor
**2/*10*	0	0	1	0	0	0	Likely Poor
**2/*2*	3	5	4	4	3	4	Poor
**2/*35*	1	0	1	0	2	1	Poor
**35/*35*	0	1	0	0	0	0	Poor
**1/*39*	0	4	0	0	1	3	indeterminate
**2/*39*	0	0	0	0	0	1	Indeterminate
**17/*39*	0	0	0	0	1	0	Indeterminate

*Metabolizer status was determined based on the Clinical Pharmacogenetics Implementation Consortium Guideline for *CYP2C19* genotype and clopidogrel therapy.

SEB: Southeastern Bantu; YRI: Yoruba in Ibadan; MSL: Mende in Sierra Leone; LWK: Luhya in Webuye, Kenya; GWD: Mandinka in western divisions of The Gambia; ESN: Esan in Nigeria.

Supplementary Table 3 : Coriell IDs for participants with publicly available high coverage WGS data containing potential novel *CYP2C19* star alleles. The Coriell DNA samples from these participants could potentially be added to GeT-RM reference materials for *CYP2C19* genotyping in pharmacogenomics testing laboratories. The *CYP2C19* variant positions in this table are according to RefSeq gene (NG_008384.3) and the protein positions are according to NP_000760.1.

Coriell IDs	Predicted diplotype	Population	Project
SSA predicted diplotypes			
HG03091, HG03225, HG03437, HG03485	**1*/[**1* + rs146991374~95056~A>T (K432I)]	MSL	1000G
HG03518	**1*/[**1* + rs150152656~ 17767~C>T (T130M)]	ESN	1000G
HG03271	**30*/[**1* + rs145328984~17426~C>T (R73C)]	ESN	1000G
HG03445	**1*/[**1* + rs201306972~ 17427~G>A (R73H)]	MSL	1000G
HG03212, HG03449, NA19314, NA19327, NA19380, NA19404, NA19455	**17*/[**2* + rs145119820~ 17715~G>A (V113I)	MSL, LWK, GWD	1000G
NA19452	[**1/*2*] + rs183701923~ 22893~C>T (R186C)	LWK	1000G
NA19403	[**17/*35*] + rs192154563~ 95085~C>T (R442C)	LWK	1000G
HG02979	[**2/*2*] + rs201306972~ 17427~G>A (R73H) +rs537050749~ 24271~A>G (N258S)	ESN	1000G
HG03428	[**1/*17*] + rs375760063~ 17856~A>G (K160E)	MSL	1000G
HG02953	[**1/*17*] + rs545642100~ 17526~C>G (A106G)	ESN	1000G
NA19307	[**2/*17*] + rs527499296~ 92332~A>C (E415D)	LWK	1000G
NA19395	[**1/*3*] + rs144056033~ 92303~A>G (M406V)	LWK	1000G
Other global predicted diplotypes			
HG04141, NA20897, NA20882	**2*/ [**38* +rs577255883~ 24282~G>A] (D262N) **1*/[**38*+rs577255883~24282~G>A] (D262N)	SAS	1000G
HG04033,HG03685,NA20885, HG03778, HG03685, HG03714,	**2*/ [**2* +rs61311738~ 22855~C>T] (A173V) [**2* +rs61311738~ 22855~C>T] (A173V)/ [*2+rs556994963~ 24276~C>A] (P260T) [*2+ rs61311738~ 22855~C>T] (A173V)/ [*2+ rs142974781~ 17826~C>T] (R150C)	SAS	1000G

HG01272, HG01512 HG01465	*1/ [*38* +rs149590953~ 17772~C>T] (R132W) *17/ [*38* +rs149590953~ 17772~C>T] (R132W)	AMR, EUR	1000G
HG03680, HG03887	*1/[*1*+rs572853437~ 5189~C>G] (T55S) *17/[*1*+rs572853437~ 5189~C>G] (T55S)	EUR	1000G
HG01170	*1/ [*1*+rs186489608~ 22930~T>C] (M198T)	AMR	1000G
HG04195	*1/ [*1*+rs542090374~ 95226~C>T] (P289S)	SAS	1000G
NA20412	*1/ [*1*+rs574458036~ 24236~G>T] (E246D)	African American/ Afro-Caribbean	1000G
HG00138	*1/ [*1*+rs201306972~ 17427~G>A] (R73H)	EUR	1000G
HG02275	*1/ [*1*+rs559087813~ 5149~A>G] (I42V)	AMR	1000G
HG02521	*1/ [*1*+rs559628884~ 62827~C>A] (N277K)	EAS	1000G
HG02081, HG02064	*1/ [*1*+rs28399505~ 17430~T>C] (M74T)	EAS	1000G
HG03600, NA21104	*1/ [*34*+rs3758581~ 85186~A>G] (I331V)	SAS	1000G
NA20761	[*1/*17*] + rs188851578~ 92237~G>C (G384R)	EUR	1000G
HG03951	[*1/*17*] +rs540369401~ 95091~G>C (E444Q)	SAS	1000G
NA20828	[*1/*17*] +rs201132803~ 85229~T>A (M345K)	EUR	1000G
HG01365	[*1/*17*] +rs183701923~ 22893~C>T (R186C)	African American/ Afro-Caribbean	1000G
HG00160	[*1/*17*] +rs200150287~ 17751~C>T (R125C)	EUR	1000G
HG01257, HG01479	[*1/ *17*] +rs200347843~ 17535~G>C (G109A)	African American/ Afro-Caribbean	1000G
HG04002	[*1/*2*] +rs368758960~ 85198~C>T (R335W)	EUR	1000G
HG01970	[*1/*2*] +rs144036596~ 85273~G>A] (D360N)	African American/ Afro-Caribbean	1000G
HG00638	[*1/*2*] +rs145119820~ 17715~G>A (V113I)	AMR	1000G
HG04164	[*1/*38*] +rs547822797~ 62833~G>T (Q279H)	SAS	1000G
HG03985	[*1/*38*] +rs59734894~ 85180~C>T (R329C)	SAS	1000G

SEB: Southeastern Bantu; MSL: Mende in Sierra Leone; LWK: Luhya in Webuye, Kenya; GWD: Mandinka in western divisions of The Gambia; ESN: Esan in Nigeria; AMR: admixed American; SAS: South Asian; EAS: East Asian; EUR: European; 1000G: 1000 Genomes Project; AWI-Gen: Africa Wits-INDEPTH Partnership for Genomic Research; GeT-RM: Genetic Testing Reference Material Coordination Program.

Supplementary Table 4: ADME-optimised variant annotations for *CYP2C19* allele-defining variants in this study**.**

#	Variant ID	Consequence	CADD	SIFT	PolyPhen-2	LRT	PROVEAN	VEST4	LOFTEE*	Consensus
1	rs6413438	Missense (P227L) (**10*)	X	X	X	X	X	X		X
2	rs146991374	Missense (K432I)	X	X	X	X	X	X		X
3	rs183701923	Missense (R186C)	X	X	X	X	X			X
4	rs17884712	Missense (R144H) (**9*)	X	X	X	X		X		X
5	rs150152656	Missense (T130M)	X	X	X	X				X
6	rs192154563	Missense (R442C)	X	X		X	X			X
7	rs145328984	Missense (R73C) (**30*)	X	X	X					X
8	rs4986893	Stop gained (W212*) (**3*)	X					X	X	X
9	rs17885179	Missense (E122A) (**39*)	X			X				
10	rs201306972	Missense (R73H)		X						
11	rs17878459	Missense (E92D)				X				
12	rs537050749	Missense (N258S)				X				
13	rs17879685	Missense (R410C) (**13*)		X						
14	rs17882687	Missense (I19L) (**15*)								
15	rs545642100	Missense (A106G)								
16	rs145119820	Missense (V113I)								
17	rs375760063	Missense (K160E)								
18	rs3758581	Missense (I331V) (*1)								
19	rs144056033	Missense (M406V)								
20	rs527499296	Missense (E415D)								
21	rs12769205	Splice defect (**2,*35)*								
22	rs4244285	Synonymous variant (P227P)								
23	rs55640102	Stop lost (*491C) (**12*)								
24	rs112197069	Missense (F110Y)								
25	rs12248560	Expression (**17*)								
26	rs550527959	Missense (D360V)	X	X	X	X	X	X		X
27	rs140278421	Missense (R186P) (**22*)	X	X	X	X	X	X		X
28	rs577255883	Missense (D262N)	X	X	X	X	X	X		X
29	rs201132803	Missense (M345K)	X	X	X	X	X	X		X
30	rs562912432	Missense (I387T)		X	X	X	X	X		X
31	rs540369401	Missense (E444Q)	X	X	X	X		X		X
32	rs41291556	Missense (W120R) (**8*)	X	X	X	X		X		X
33	rs188851578	Missense (G384R)	X	X	X		X	X		X
34	rs556994963	Missense P260T)	X	X	X	X	X			X
35	rs368758960	Missense (R335W)	X	X	X		X			X
36	rs186489608	Missense (M198T)	X	X	X		X			X
37	rs200150287	Missense (R125C)	X	X	X	X				X
38	rs149590953	Missense (R132Q) (**6*)	X	X	X	X				X
39	rs542090374	Missense (P489S)		X	X	X	X			X
40	rs61311738	Missense (A173V)	X		X			X		X
41	rs144036596	Missense (D360N)			X	X	X			X

42	rs28399504	Start Lost (M1V) (**4*)		X				X
43	rs181297724	Missense (A161P)	X	X				
44	rs137862636	Intron variant	X					
45	rs200347843	Missense (G109A)		X		X		
46	rs367674480	Missense (R410H)		X				
47	rs118203757	Missense (R335Q) (**24*)			X			
48	rs59734894	Missense (R329C)					X	
49	rs559628884	Missense (N277K)					X	
50	rs142974781	Missense (R150C)					X	
51	rs574458036	Missense (E246D)						
52	rs559087813	Missense (I42V)						
53	rs547822797	Missense (Q279H)						
54	rs28399505	Missense (M74T)						

*LOFTEE only provides predictions for loss of function variants

The presence of an 'X' indicates a deleterious outcome for the variant as predicted by the *in silico* tool. The ADME-optimised deleteriousness thresholds are as follows: CADD > 19.19; SIFT < 0.0376; Polyphen-2 > 0.3841; LRT < 0.0025; PROVEAN <-3.286; VEST4 > 0.4534. Variants 1-25 were present in the sub-Saharan African populations in this study while the rest were only found in the other global populations from the 1000 Genomes Project Dataset

Appendix A

Approved Research Protocol

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

CANDIDATE'S SURNAME: Booyse	FIRST NAME/S: Ross Peter	STUDENT NUMBER: 1827129
CURRENT QUALIFICATIONS: BSc Biological Science (distinction) BHSc Human Genetics (distinction)		
TEL: 011 465 0883	CELL: 082 370 7619	E-MAIL: 1827129@students.wits.ac.za / FAX: N/A
DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED: MSc Genomic Medicine		
PART-TIME OR FULL-TIME: FULL-TIME		
FIRST REGISTERED FOR THIS DEGREE:	TERM : 1	YEAR: 1
DEPARTMENT: Human Genetics		
TITLE OF PROPOSED RESEARCH: Characterisation of variation in the *CYP2C19 gene* in African populations		
CANDIDATE'S SIGNATURE:		DATE: 14 June 2022
SUPERVISOR 1 (NAME & SURNAME): Mr. David Twesigomwe		% Supervision 50%
SUPERVISOR'S QUALIFICATIONS: BSc; PhD (awaiting examiners)		
SUPERVISOR'S DEPARTMENT: Division of Human Genetics, School of Pathology		
SUPERVISOR'S ADDRESS / TEL / E-MAIL: 1989491@students.wits.ac.za		
SUPERVISOR 2 (NAME & SURNAME): Prof. Scott Hazelhurst		% Supervision 50%
SUPERVISOR'S QUALIFICATIONS: Professor; PhD		
SUPERVISOR'S DEPARTMENT: School of Electrical and Information Engineering		
SUPERVISOR'S ADDRESS / TEL / E-MAIL: scott.hazelhurst@wits.ac.za		
SYNOPSIS OF RESEARCH: see next page		

WITS ETHICS NOT REQUIRED: WITS ETHICS PENDING: WITS ETHICS APPROVED: (circle appropriate symbol) *	Yes No Yes No Yes No	IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE: M220293
*Please note the final human ethics clearance certificate or animal ethics certificate must be available prior to starting research		
As supervisor/s, I/we confirm that I have read the protocol which has been submitted for assessment.		
SIGNATURE OF SUPERVISOR/S:	.....	.....
SIGNATURE PG OFFICE STAFF	REGISTERED YES..✓.. NO.....	STAMP

SYNOPSIS OF RESEARCH

Background and aim: The implementation of pharmacogenomics (PGx) in clinical settings has the potential to improve drug safety and efficacy across Africa and worldwide. However, this requires characterisation of variation in genes that play a role in absorption, distribution, metabolism and excretion(ADME) of drugs. Of the ADME genes *CYP2C* genes (*CYP2C9, CYP2C19, CYP2C8* and *CYP2C18*) are important for various important drug responses and are listed as priority in context of the disease burden inAfrica. However, the current catalogues of *CYP2C* pharmacogenetic variation is missing critical information on African-specific variants and their distribution across diverse African populations. This is particularly the case for *CYP2C19* for which the focus of this study will be on. Therefore, this study aims to characterise the distribution of pharmacogenomically relevant alleles in the *CYP2C19* gene across diverse African populations. **Methods**: The datasets in this study include high coverage genomes from five African populations represented in the 1000 Genomes Project repository. In addition, this study will analyze 100 genomes from south-eastern Bantu speakers from South Africa collected as part of the AWI- Gen dataset. *CYP2C19* gene haplotypes and diplotypes will be called using the StellarPGx pipeline, and all variants will be annotated using plug-ins within Ensembl's variant effect predictor (VEP). Potential novel African-specific alleles predicted by StellarPGx will be manually curated to avoid any artefacts.

Expected results: We expect to characterise the distribution of both common and rare alleles in *CYP2C19* across diverse African populations, to identify potential novel *CYP2C19* haplotypes across African populations and to annotate novel allele-defining variants in *CYP2C19* in comparison with existing core variants.

Conclusion: Characterisation of African-specific variation in the *CYP2C19* gene could help greatly improve drug efficacy and response of various important FDA approved medications such as clopidogrel. Furthermore, characterisation of novel variants and star alleles in diverse African populations could aid in guiding development of gene panels specific for these population to support the relief of the disease burden in Africa.

Characterisation of variation in the *CYP2C19* gene in African populations

Research Project Proposal

Student: Ross Booyse

Student number: 1827129

To the Faculty of Health Sciences, University of the Witwatersrand, in partial fulfilment of the requirements for the Degree of

Masters in Genomic Medicine

Submitted: 14 June 2022

Supervisor: David Twesigomwe

Co-supervisor: Scott Hazelhurst

Table of contents

Table of Figures

1. Introduction

Individuals respond differently to clinically prescribed medications, in part due to genetic variation. This response ranges from treatment non-response to severe adverse drug reactions (ADRs) (Zhao et al., 2021). The implementation of pharmacogenomics (PGx) in clinical settings has the potential to improve treatment safety and efficacy across Africa and in clinical settings worldwide (Desta et al., 2019; Lee et al., 2022). However, this requires comprehensive characterisation of variation in genes that play a role in absorption, distribution, metabolism and excretion (ADME) of drugs (Arbitrio et al., 2018).

1.1 Pharmacogenomic significance of *CYP2C19*

Of the ADME genes, cytochrome P450 (CYP) genes are widely studied as they encode enzymes responsible for 80% of all drug metabolism and 50% of drug elimination (Zhao et al., 2021). In particular, the *CYP2C19* gene, which is found on the *CYP2C* gene cluster alongside *CYP2C8*, *CYP2C9* and *CYP2C18* on chromosome 10 (Figure 1), is important especially in the metabolism of anti-depressant and antiplatelet medication (Botton et al., 2021). *CYP2C19* genotype-guided therapy is increasingly being adopted in clinical practice (Aldrich et al., 2019). The phenotypes associated with *CYP2C19* ranges from poor metaboliser (PM) to ultra-rapid metaboliser (UM) which can be explained by certain variation within this gene.

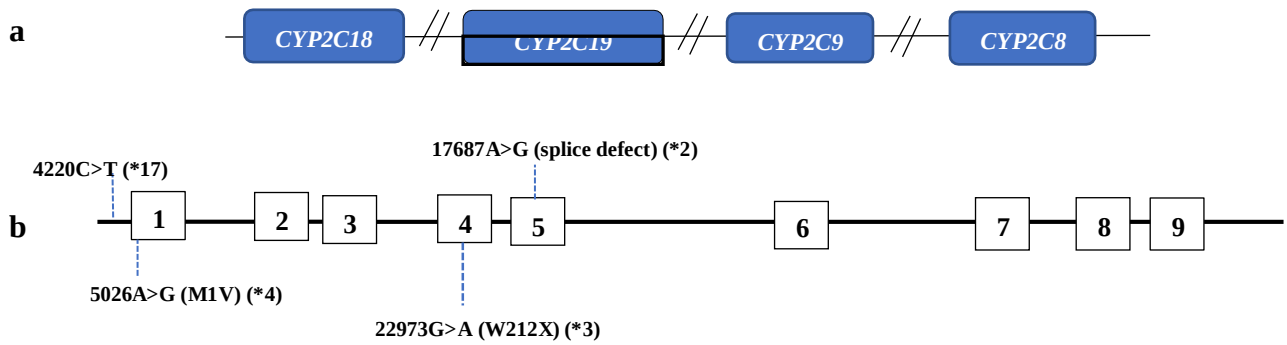

Figure 1: Overview of the *CYP2C* gene locus. (**a**) Graphical summary of location of genes on the *CYP2C* cluster on chromosome 10. This includes *CYP2C18*, *CYP2C19*, *CYP2C9* and *CYP2C8*. Numbered boxes represent exons. (**b**) Core variants which define *CYP2C19*2* to **4* and **17*.

In the context of the disease burden and common treatments in Africa, *CYP2C19* was suggested as part of the priority pharmacogenes to investigate in African populations by Alessandrini and Pepper (2014). However, the current catalogues of important *CYP2C19* pharmacogenetic variation are missing critical information on African-specific variants and their distribution across diverse African populations (see https://www.pharmvar.org/gene/CYP2C19). Therefore, there is a need for characterisation of important variation in the *CYP2C19* gene in understudied African populations to inform drug dosing guidelines for clopidogrel, antidepressants and other drugs metabolised by CYP2C19.

1.2 *CYP2C19* genomic architecture and star alleles

The *CYP2C19* gene consists of nine exons which cover 93.9 kb on chromosome 10q23 . This makes it the largest of the genes in the *CYP2C* gene cluster (Figure 1). The Pharmacogene Variation Consortium (PharmVar) currently catalogues 39 star (*) alleles (i.e. haplotypes) for *CYP2C19* defined by various combinations of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), and/or structural variations (Botton et al., 2019). From these star alleles *CYP2C19*2, *3* and **4* are the most extensively studied. Although all of these are associated with the PM phenotype (depending on the diplotype), *CYP2C19*2* has the highest correlation with PMs in African populations (Alessandrini & Pepper, 2014). This star allele has a wide range of frequencies (7-23%) across African populations (including populations from South Africa, Egypt, Kenya, Ghana and Ethiopia, to name a few) according to a paper by Alessandrini, et al., (2013). In the same paper it is also noted that this frequency is lower than that observed in Asian populations (23-38%). In an article by Dodgen et al., (2015), it was observed that the *CYP2C19*2* allele frequency was approximately 2% in their South African cohort, however the *CYP2C19*3* allele was not present at all in these individuals.

CYP2C19 also harbours important star alleles that are associated with the UM phenotype. However, this phenotype is mainly correlated with the *CYP2C19*17* allele. This star allele has an observed frequency of 17% in Sub-Saharan African (SSA) populations, which is significantly higher than that seen in East Asians (2%) (Botton et al., 2021). The frequency in SSA is comparable to that in a paper by Drögemöller et al., (2010), which investigated South African Xhosa (13%) and Cape Mixed Ancestry populations (14%). Failure to account for African-specific star alleles such as *CYP2C19*17* in gene panels or clinical guidelines would therefore have considerable consequences for clopidogrel and anti-depressant response in African individuals.

1.3 Mining high coverage data for variation in the *CYP2C19* in African populations

One of the challenges with *CYP2C19* PGx studies is the ability to detect both rare and common variants as well as structural variations in the population (Bienfait et al., 2021). Most pharmacogenetics testing platforms used today do not include rare variation, and rather focus on locus-specific methods designed to identify known variants (Gaedigk et al., 2022). This could be problematic during clinical pharmacogenomic implementation for African populations since much of their variation is not yet characterised. Another challenge for PGx in Africa lies in the fact that most of the previously available African datasets were low coverage data (Gurdasani et al., 2014) and/or whole exome sequences (Auton et al., 2015). The low coverage data hinders the ability to fully characterise novel and/or rare African-specific star alleles (Pereira et al., 2021), and also structural variations.

The recent increase in the availability of high coverage whole genome sequence (WGS) data representative of African populations (Byrska-Bishop et al., 2021; Choudhury et al., 2020) presents the opportunity to investigate the distribution of common and African-specific *CYP2C19* variation. In addition to this, recent development of tools that aid in the characterisation of star alleles from high throughput sequence (HTS) data, allows for a scalable analysis of the data (Lee et al., 2019; Twesigomwe et al., 2021). The importance of advances in next generation sequencing (NGS) data generation and bioinformatic analysis to addressing the paucity of information relevant to implementing precision medicine strategies involving *CYP2C19* drug-gene pairs across Africa cannot be overstated (Schwarz et al., 2019).

1.4 Study aim

Therefore, this study aims to characterise the distribution of pharmacogenomically relevant alleles in the *CYP2C19* gene across diverse African populations.

1.5 Objectives

1) To characterise the distribution of both common and rare alleles in *CYP2C19* across diverse African populations.

2) To identify potential novel *CYP2C19* haplotypes across African populations.

3) To annotate novel allele-defining variants in *CYP2C19* in comparison with existing core variants.

2. Methodology

2.1 Study design

This is a comparative study involving secondary analysis of genomic data from diverse African populations.

2.1.1 Data sources

The latest high coverage (30x) public data release from the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2021) contains 504 African whole genome sequence (WGS) genomes (N=504) which are all included in this study (Figure 2). In addition to this data, we will perform secondary analysis on the WGS dataset from the Africa Wits-INDEPTH partnership for Genomics studies (AWI-Gen) which consists of 100 high-depth (40x) genomes from south-eastern Bantu participants (Ramsay et al., 2016). The total sample size for this project is therefore 604 participants (excluding those mentioned in the eligibility criteria section below). All WGS data retrieved has already undergone quality control under their respective studies (Byrska-Bishop et al., 2021; Ramsay et al., 2016).

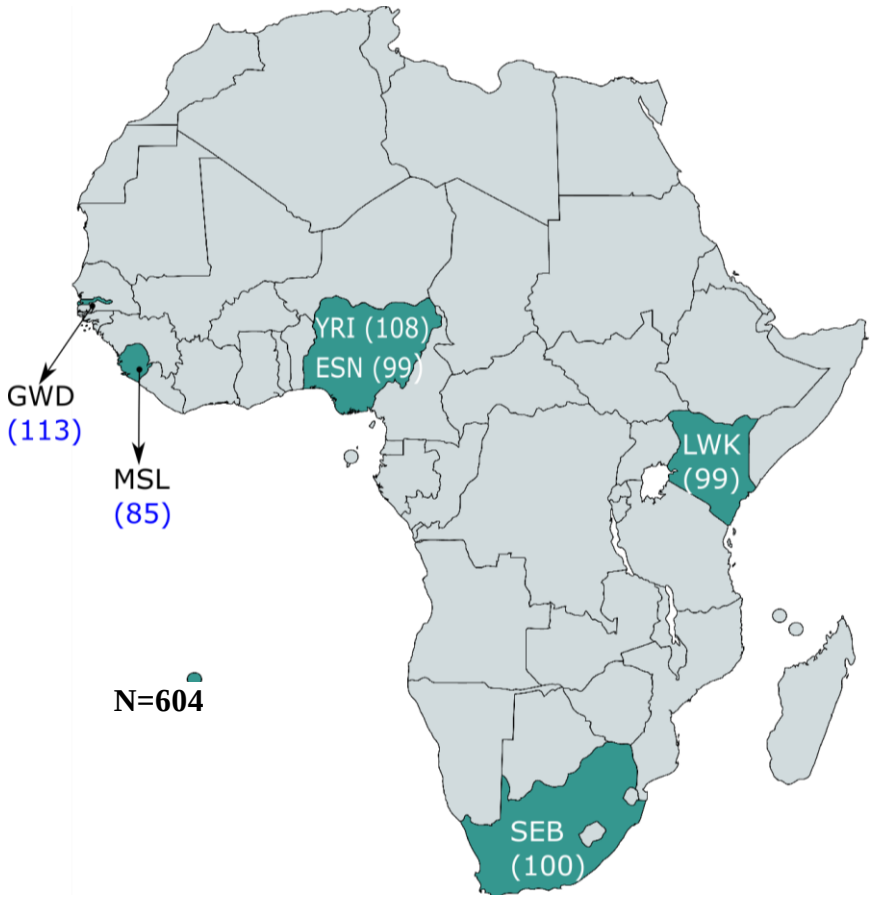

Figure 2: Map of Africa showing the breakdown of populations and their respective sample sizes. This includes individuals from the 1000 Genomes and AWI-GEN project for a total of 604 individuals. SEB, south east Bantu; LWK, Luhya in Webuye Kenya; YRI, Yoruba in Ibadan Nigeria; ESN, Esan in Nigeria; MSL, Mende in Sierra Leone; GWD, Western division of The Gambia

2.1.2 Eligibility Criteria

All continental African individuals for whom there is high coverage WGS data (~30x) released by the 1000 Genomes will be included in the study, in addition to the aforementioned WGS data from the AWI-Gen project. This excludes African Caribbeans in Barbados (ACB) and the Americans of African Ancestry in South West USA (ASW) for whom high coverage data also exists as part of the 1000 Genomes Project dataset.

2.2 Bioinformatic and statistical analysis

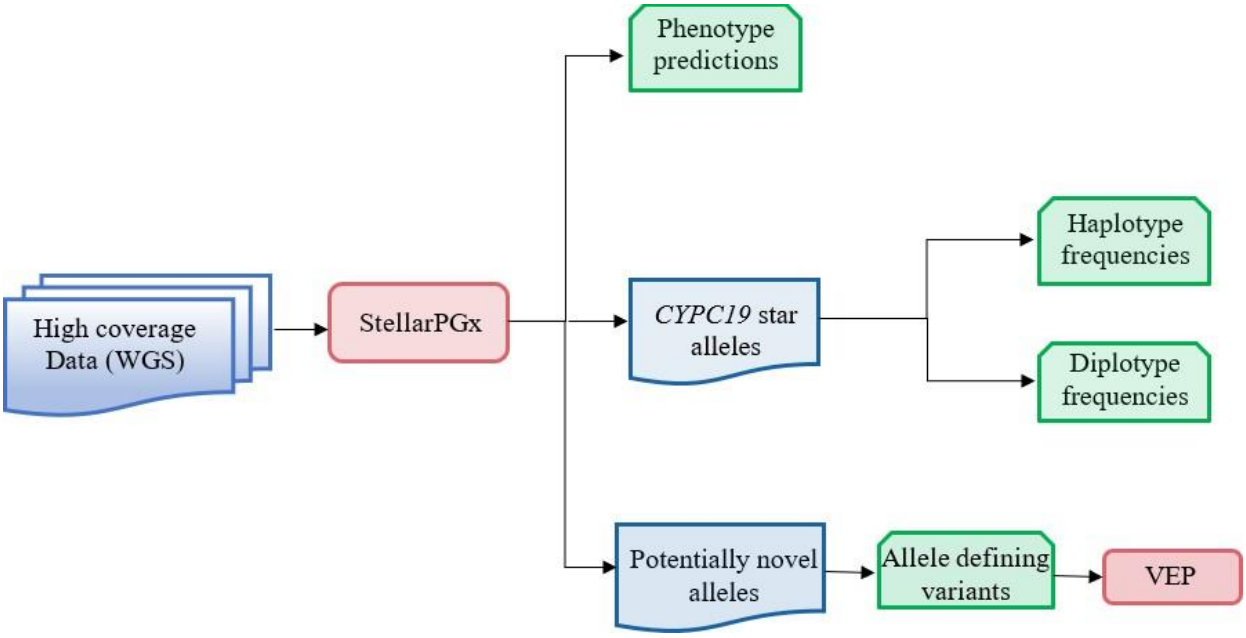

Figure 3: Workflow for bioinformatic analysis of the high coverage input data represented in the 1000 Genomes and AWI-Gen datasets. VEP, Variant Effect Predictor; WGS, Whole Genome Sequence

2.2.1 *CYP2C19* star allele and variant analysis

CYP2C19 star alleles from the 1000 Genomes and AWI-Gen datasets will be called using the Nextflow pipeline StellarPGx (Twesigomwe et al., 2021) (Figure 3). This *in silico* tool provides a scalable and reproducible approach to determining haplotype and diplotype information in *CYP* genes according to the star allele nomenclature utilised by PharmVar.

The star allele calling process is achieved via various steps within the workflow of StellarPGx (https://github.com/SBIMB/StellarPGx). The input is a sample read-alignment (BAM) file, i.e. the high coverage WGS data shown in Figure 3. StellarPGx detects single nucleotide variants (SNVs) from the WGS data using GraphTyper (Eggertsson et al., 2017). GraphTyper detects variants via a genome graph approach rather than a linear reference-based approach, thus enabling accurate variant detection as a function of improved read realignment and reduced reference allele bias.

Following genome-graph based variant detection, StellarPGx determines *CYP2C19* diplotypes by detecting candidate diplotypes (based on VCF output) and combining this information with copy number alterations detected by the tool (Twesigomwe et al., 2021). The candidate diplotypes are inferred based on whether the core variant combinations in the VCF file match allele-defining variant combinations in the StellarPGx database – which is compiled based on the PharmVar database. Core variants typically include missense variants, frameshift variants, stop gain/loss and splice site variants. In some cases, sub-allele variants need to be considered in order to resolve diplotypes that share the same combination of core variants. StellarPGx relies on read depth distributions and variant allele depth ratios when determining structural variations e.g. figuring out whether a sample with a gene deletion has **36* (full *CYP2C19* deletion) or **37* (partial *CYP2C19* deletion). For individuals with potential novel alleles, StellarPGx includes a flag in the output for extra manual inspection and highlights the core variants. The allele-defining variants will be subsequently annotated as described in section 2.2.2. For *CYP2C19*, participant phenotypes will be predicted using the activity score system recommended by the clinical pharmacogenomics implementation consortium (CPIC).

The distribution of star allele haplotype and diplotype frequencies in the different ethnolinguistic groups will be determined by using the summary output from StellarPGx. Star allele frequencies between ethnolinguistic groups will be compared by using Fisher's exact test. To add to this, any deviations from Hardy-Weinberg Equilibrium will be evaluated using the statistical programme R (https://www.r-project.org/).

2.2.2 *CYP2C19* Variant annotation

To obtain information on the consequences of the novel allele defining variants identified in *CYP2C19*, VCF files from StellarPGx will be run through Ensembl's Variant Effect Predictor (VEP) (McLaren et al., 2016) (Figure 3). The VEP plug-ins SIFT (Kumar et al., 2009),

Polyphen-2 (Adzhubei et al., 2013), CADD (Rentzsch et al., 2019), LRT (Fujita et al., 2010), PROVEAN (Choi & Chan, 2015), MutationAssessor (Reva et al., 2011) and VEST4 (Carter et al., 2013) will be used to predict deleteriousness of variants in the *CYP2C19* gene. The combination of these tools provides a more accurate prediction for variant impacts in pharmacogenes such as *CYP2C19* (Zhou et al., 2019). Functional predictions made by these tools will be following the ADME-optimised parameters suggested and validated by Zhou et al. (2019). High-impact variants will be defined as those predicted to be deleterious by at least 3 of the VEP plugins.

3. Ethics

No direct contact will be made with human participants, nor will any new data be added to any of the datasets. In accordance with the POPIA act, all patient identifiers have been removed from the WGS data and codes are used to refer to patient samples. The analysis tools (VEP and StellarPGx) to be used in this study and part of the data (1000 Genomes project datasets) are open-source and publicly available. However, the AWI-Gen data is not publicly available – approval for access to this data was granted by the data custodian Prof Michele Ramsay (AWI-Gen principal investigator). Furthermore, an application for a sub-study ethics was submitted on the 7th of March 2022 and was approved on the 16th of March 2022 (protocol number: M220293)

4. Timing

Task	**Year 1 (2022)**											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Literature review												
Ethics application												
Writing and presenting proposal document												
Allele calling and variant annotation												
Statistical analysis												
Manuscript preparation												
Submission and presentation												

5. Funding

There is no funding required for this project since it involves analysis of secondary data that is already on the Wits Core computing cluster (PI: Prof Scott Hazelhurst), and all the *in silico* tools being used are publicly available and free of charge.

6. Limitations

The use of short read technology utilised for WGS sequencing in the 1000 Genomes and AWI-Gen dataset presents a limitation to this project. This is because the *CYP2C* gene cluster, namely *CYP2C19* and *CYP2C9*, are known to harbour structural variations due to the high homology between the two genes. Short read technology has limited success in enabling accurate analysis of genomic regions with repetitive sequences and complex structural variants, thus it may result in unreliable diplotype calls for individuals with complex novel alleles. Key variations in these regions could therefore by missed or misinterpreted by this approach. In addition, molecular functional assays, such as structural bioinformatics, are not in the scope of this study. This presents challenges when providing accurate variant annotations for key variants as *in silico* predictions do not provide information on whether the potential effect of some deleterious variants are actually impacting drug-gene pair associations.

7. References

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, Chapter 7, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

Aldrich, S. L., Poweleit, E. A., Prows, C. A., Martin, L. J., Strawn, J. R., & Ramsey, L. B. (2019). Influence of CYP2C19 Metabolizer Status on Escitalopram/Citalopram Tolerability and Response in Youth With Anxiety and Depressive Disorders . *Frontiers in Pharmacology, 10(1)*. https://doi.org/10.3389/fphar.2019.00099

Alessandrini, M., & Pepper, M. S. (2014). Priority pharmacogenetics for the African continent: focus on CYP450. *Pharmacogenomics*, *15*(3), 385–400. https://doi.org/10.2217/PGS.13.252

Arbitrio, M., Di Martino, M. T., Scionti, F., Barbieri, V., Pensabene, L., & Tagliaferri, P. (2018). Pharmacogenomic Profiling of ADME Gene Variants: Current Challenges and Validation Perspectives. *High-Throughput 2018, 7*(4), 40. https://doi.org/10.3390/HT7040040

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., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature* , *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bienfait, K., Chhibber, A., Marshall, J. C., Armstrong, M., Cox, C., Shaw, P. M., & Paulding, C. (2021). Current challenges and opportunities for pharmacogenomics: perspective of the Industry Pharmacogenomics Working Group (I-PWG). *Human Genetics*, 1–9. https://doi.org/10.1007/S00439-021-02282-3/TABLES/1

Botton, M. R., Lu, X., Zhao, G., Repnikova, E., Seki, Y., Gaedigk, A., Schadt, E. E., Edelmann, L., & Scott, S. A. (2019). Structural variation at the CYP2C locus:Characterization of deletion and duplication alleles. *Human Mutation*, *40*(11), 37. https://doi.org/10.1002/HUMU.23855

Botton, M. R., Whirl-Carrillo, M., Del Tredici, A. L., Sangkuhl, K., Cavallari, L. H., Agúndez, J. A. G., Duconge, J., Lee, M. T. M., Woodahl, E. L., Claudio-Campos, K., Daly, A. K., Klein, T. E., Pratt, V. M., Scott, S. A., & Gaedigk, A. (2021). PharmVar GeneFocus: CYP2C19. *Clinical Pharmacology & Therapeutics*, *109*(2), 352–366. https://doi.org/10.1002/CPT.1973

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy-Gallego, E., Consortium, T. H. G. S. V., Flicek, P., Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *BioRxiv*. https://doi.org/10.1101/2021.02.06.430068

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, *14* (Suppl 3). https://doi.org/10.1186/1471-2164-14-s3-s3

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*(16), 2745–2747. https://doi.org/10.1093/bioinformatics/btv195

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y. J., Ghoorah, A. W., Dareng, E., Odia, T., Falola, O., Adebiyi, E., Hazelhurst, S., Mazandu, G., Nyangiri, O. A., Mbiyavanga, M., … Hanchard, N. A. (2020). High-depth African genomes inform human migration and health. *Nature*, *586*(7831), 741–748. https://doi.org/10.1038/s41586-020-2859-7

Desta, Z., Gammal, R. S., Gong, L., Whirl-Carrillo, M., Gaur, A. H., Sukasem, C., Hockings, J., Myers, A., Swart, M., Tyndale, R. F., Masimirembwa, C., Iwuchukwu, O. F., Chirwa, S., Lennox, J., Gaedigk, A., Klein, T. E., & Haas, D. W. (2019). Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2B6 and Efavirenz-Containing Antiretroviral Therapy. *Clinical Pharmacology and Therapeutics, 106*(4), 726-733. https://doi.org/10.1002/cpt.1477

Dodgen, T. M., Drögemöller, B. I., Wright, G. E. B., Warnich, L., Steffens, F. E., Cromarty, A. D., Alessandrini, M., & Pepper, M. S. (2015). Evaluation of predictive CYP2C19 genotyping assays relative to measured phenotype in a South African cohort. *Pharmacogenomics*, *16*(12), 1343–1354. https://doi.org/10.2217/PGS.15.80

Drögemöller, B. I., Wright, G. E. B., Niehaus, D. J. H., Koen, L., Malan, S., Da Silva, D. M., Hillermann-Rebello, R., La Grange, A. M., Venter, M., & Warnich, L. (2010). Characterization of the genetic profile of CYP2C19 in two South African populations. *Pharmacogenomics*, *11*(8), 1095–1103. https://doi.org/10.2217/PGS.10.90

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics, 49,* 1654–1660. https://doi.org/10.1038/ng.3964

Fujita, A., Kojima, K., Patriota, A. G., Sato, J. R., Severino, P., & Miyano, S. (2010). A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets. *Bionformatics Applications Note*, *26*(18), 2349–2351. https://doi.org/10.1093/bioinformatics/btq427

Gaedigk, A., Boone, E. C., Scherer, S. E., Lee, S., Numanagić, I., Sahinalp, C., Smith, J. D., McGee, S., Radhakrishnan, A., Qin, X., Wang, W. Y., Farrow, E. G., Gonzaludo, N., Halpern, A. L., Nickerson, D. A., Miller, N. A., Pratt, V. M., & Kalman, L. V. (2022). CYP2C8, CYP2C9 and CYP2C19 characterization using Next Generation Sequencing and Haplotype Analysis: A GeT-RM Collaborative Project. *The Journal of Molecular Diagnostics, 24*(4), 337-350. https://doi.org/10.1016/J.JMOLDX.2021.12.011

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2014). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332. https://doi.org/10.1038/nature13997

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1082. https://doi.org/10.1038/nprot.2009.86

Lee, C. R., Luzum, J. A., Sangkuhl, K., Gammal, R. S., Sabatine, M. S., Stein, C. M., Kisor, D. F., Limdi, N. A., Lee, Y. M., Scott, S. A., Hulot, J. S., Roden, D. M., Gaedigk, A., Caudle, K. E., Klein, T. E., Johnson, J. A., & Shuldiner, A. R. (2022). Clinical Pharmacogenetics Implementation Consortium Guideline for *CYP2C19* Genotype and Clopidogrel Therapy: 2022 Update. *Clinical Pharmacology and Therapeutics*. https://doi.org/10.1002/CPT.2526

Lee, S., Wheeler, M. M., Patterson, K., McGee, S., Dalton, R., Woodahl, E. L., Gaedigk, A., Thummel, K. E., & Nickerson, D. A. (2019). Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genetics in Medicine*, *21*(2), 361–372. https://doi.org/10.1038/s41436-018-0054-0

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Pereira, L., Mutesa, L., Tindana, P., & Ramsay, M. (2021). African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics*, *22*(5), 284–306. https://doi.org/10.1038/s41576-020-00306-8

Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., Gómez-Olivé, X., Jaff, N., Sorgho, H., Wagner, R., Khayeka-Wandabwa, C., Choudhury, A., Hazelhurst, S., Kahn, K., Lombard, Z., Mukomana, F., Soo, C., Soodyall, H., Wade, A., … Sankoh, O. (2016). H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global Health, Epidemiology and Genomics*, *1*, 1–13. https://doi.org/10.1017/GHEG.2016.17

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), 118. https://doi.org/10.1093/nar/gkr407

Schwarz, U. I., Gulilat, M., & Kim, R. B. (2019). The Role of Next-Generation Sequencing in Pharmacogenetics and Pharmacogenomics. *Cold Spring Harbor Perspectives in*

Medicine, *9*(2). https://doi.org/10.1101/CSHPERSPECT.A033027

Twesigomwe, D., Drögemöller, B. I., Wright, G. E. B., Siddiqui, A., da Rocha, J., Lombard, Z., & Hazelhurst, S. (2021). StellarPGx: A Nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clinical Pharmacology & Therapeutics*, *110*(3), 741-749. https://doi.org/10.1002/cpt.2173

Zhao, M., Ma, J., Li, M., Zhang, Y., Jiang, B., Zhao, X., Huai, C., Shen, L., Zhang, N., He, L., & Qin, S. (2021). Cytochrome P450 Enzymes and Drug Metabolism in Humans. *International Journal of Molecular Sciences*, *22*(23), 12808. https://doi.org/10.3390/IJMS222312808

Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M., & Lauschke, V. M. (2019). An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics Journal*, *19*(2), 115–126. https://doi.org/10.1038/s41397-018-0044-2

Appendix B

Ethics Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO:	Mr R Booyse Faculty of Health Sciences Sydney Brenner Institute for Molecular Bioscience and Division of Human Genetics Medical School University E-mail: 1827129@students.wits.ac.za
CC:	Supervisor: Mr D Twesigomwe and Professor S Hazelhurst <1989491@students.wits.ac.za> and <HREC-Medical Research Office@wits.ac.za>
FROM:	Mr Iain Burns Human Research Ethics Committee (Medical) Tel: 011 717 1252 E-mail: Iain.Burns@wits.ac.za
DATE:	2022/03/16
REF:	R14/49
PROTOCOL NO:	**M220293** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)
PROJECT TITLE:	*Characterization of variation in the CYP2C gene cluster in African populations*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.

MSWorks2000/Iain0007/Clearscan wps

UNIVERSITY OF THE
WITWATERSRAND
JOHANNESBURG

R49 Mr R Booyse

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M220293

NAME: (Principal Investigator)	Mr R Booyse
DEPARTMENT:	Faculty of Health Sciences Sydney Brenner Institute for Molecular Bioscience and Division of Human Genetics Medical School University
PROJECT TITLE:	*Characterization of variation in the CYP2C gene cluster in African populations*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved conditionally
CONDITIONS:	Sub-study under M180535 The HREC requires evidence of Faculty protocol approval
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Mr D Twesigomwe and Professor S Hazelhurst
APPROVED BY:	C B Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2022/03/16

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **February** and therefore reports and re-certification will be due in the month of **February** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

22 March 2022
Date

Appendix C

Turn-it-in Report

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

Plagiarism Declaration

I Ross Peter Booyse (Student number:) am a student registered for the degree of MSc (Med) Genomic Medicine in the academic year 2022.

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" software indicating the level of plagiarism in my research document.

Signature: 

Date: 14 December 2022

final thesis for turnitin-3.docx

ORIGINALITY REPORT

7%	3%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ross P Booyse, David Twesigomwe, Scott Hazelhurst. " Characterization of haplotype distribution in African populations and comparison with other global populations ", Pharmacogenomics, 2022 Publication	1%
2	Submitted to University of Witwatersrand Student Paper	1%
3	David Twesigomwe, Britt I. Drögemöller, Galen E.B. Wright, Clement Adebamowo et al. " Characterization of Pharmacogenetic Variation in African Populations ", Clinical Pharmacology & Therapeutics, 2022 Publication	1%
4	Submitted to University of Portsmouth Student Paper	1%
5	www.ncbi.nlm.nih.gov Internet Source	1%
6	David Twesigomwe, Britt I. Drögemöller, Galen E. B. Wright, Clement Adebamowo et al.	<1%

" Characterisation of pharmacogenetic variation in African populations ", Clinical Pharmacology & Therapeutics, 2022
Publication

7	www.mdpi.com Internet Source	<1%
8	David Twesigomwe, Galen E. B. Wright, Britt I. Drögemöller, Jorge da Rocha, Zané Lombard, Scott Hazelhurst. "A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping", npj Genomic Medicine, 2020 Publication	<1%
9	www.frontiersin.org Internet Source	<1%
10	Mariamena Arbitrio, Francesca Scionti, Maria Teresa Di Martino, Licia Pensabene, Pierfrancesco Tassone, Pierosandro Tagliaferri. "Pharmacogenetics/Pharmacogenomics of Drug-Metabolizing Enzymes and Transporters", Elsevier BV, 2021 Publication	<1%
11	eprints.gla.ac.uk Internet Source	<1%
12	molecularcasestudies.cshlp.org Internet Source	<1%

13	www.scilit.net Internet Source	<1%
14	zenodo.org Internet Source	<1%
15	DeCandia, Alexandra L.. "An Inclusive Understanding of Molecular Variation in Population Genetics and Wildlife Disease Ecology.", Princeton University, 2020 Publication	<1%
16	Sherry-Ann Brown, Naveen Pereira. "Pharmacogenomic Impact of CYP2C19 Variation on Clopidogrel Therapy in Precision Cardiovascular Medicine", Journal of Personalized Medicine, 2018 Publication	<1%
17	Andrea Gaedigk. " Complexities of gene analysis and interpretation ", International Review of Psychiatry, 2013 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Appendix D

Author Guidelines for *Pharmacogenomics*

Research Article

Word limit: 5000–8000 words (excluding abstract, summary points, references and figure/table legends) Research Articles should present novel work that makes a significant impact within the scope of the journal, and which represents an important advancement in knowledge or understanding. Routine or incremental work is not suitable for full Research Articles. Research should be reported succinctly; the inclusion of detailed background discussion is to be avoided. Supporting data or further experimental details can be submitted as Supplementary Information. Articles submitted as Research Articles may be recommended to be considered as a Preliminary/Short Communication at the discretion of the Journal Editor.

Required sections for Research Articles, and Preliminary and Short Communications (for a more detailed description see Article sections):

▪ Title (maximum 120 characters)

▪ Author(s) names & affiliations

▪ Structured abstract (maximum 120 words)

▪ Plain language summary (optional; maximum 250 words)

▪ Tweetable abstract (optional; ~200 characters)

▪ Keywords (5–10)

▪ Introduction

o Should only cite directly pertinent references

o Should not include data of conclusions from the work being reported

▪ Patients & methods/Materials & methods

o Where an organization was paid or otherwise contracted to help conduct the research (e.g., data collection and management), this should be detailed

o Should include information indicating that the research was approved or exempted from the need for review by the responsible review committee (institutional or national). Where no formal ethics committee is available, a statement indicating that the research was conducted according to the principles of the Declaration of Helsinki should be included

o Information on the selection and description of participants should define how

authors measured race or ethnicity and justify their relevance

▪ Results

o Numeric results should be given not only as derivatives (e.g., percentages) but also as the absolute numbers from which the derivatives were calculated

o Statistical significance of results should be specified, if any

▪ Discussion

o Authors should distinguish between clinical and statistical significance, and avoid making statements on economic benefits and costs unless the manuscript includes the appropriate economic data and analyses

o Authors should avoid claiming priority or alluding to work that has not been completed

▪ Conclusions

▪ Summary points

▪ References

▪ Reference annotations

▪ Author contributions: brief summary of the contribution of each individual meeting the criteria to be listed as an author on the manuscript

▪ Acknowledgements: author acknowledgements, plus, where relevant, details of individuals who contributed to the article, but who did not fulfill the criteria to be listed as authors

▪ Disclosures: to include funding information, financial and/or conflict-of-interest disclosures, disclosure of any writing assistance (and the funding source for this), and any other relevant information

▪ Ethical conduct of research statement

▪ Data sharing statement (for studies reporting the original results of a clinical trial or the secondary analysis of clinical trial data only)

▪ Figures/Tables

o There are no strict limits on the number allowed (we prefer these to be included as needed to represent the data appropriately); however, in some cases where there is a high number of figures/tables (generally >8 Figures/Tables combined) the Editor will recommend that some are included as online-only supplementary materials

o Should be submitted as separate files

Referencing

Format

- Authors' names should appear without full stops in their initials
- Quote first six authors' names. If there are more than six, then quote the first three *et al*.
- A full stop follows the authors' names
- Journal names should be in italics, abbreviated to the standard format with full stops
- Volume number followed by issue number in brackets, followed by comma (not bold) i.e., Volume 3 issue 1 would be shown as 3(1),
- Full page number range seperated by a hypen with no spaces, followed by the year in bracksts and then a full stop

Journal example: Davies J, Martinec M, Delmar P *et al*. Comparative effectiveness from a single-arm trial and real-world data: alectinib versus ceritinib. *J. Comp. Eff. Res.* 7(9), 855-865 (2018).

Referencing preprints

The Future Science Group journals permit the citation of preprints in the reference list, but these should be clearly identified as such by including the article's web address on the preprint server. The citation should include the author names (in the same style as the journal example above), article title and the web address where the preprint can be found.

Preprint example: Korukluoglu G, Kolukirik M, Bayrakdar F *et al*. 40 minutes RT-QPCR assay for screening spike N501Y and HV69-70del mutations. Available at: https://www.biorxiv.org/content/10.1101/2021.01.26.428302v1..

Figures in your article

Figure guidelines

When including original figures in your article, it is helpful to bear the following in mind:

- **File submission:** All figures, tables and boxes should be **submitted as separate files**, not within the main manuscript document.
  - It is acceptable to include e.g., one table file with multiple tables included, but this should be a separate document to the main text file.
- **File formats:** Figures, tables and boxes should be submitted in an **editable format** where possible. Figures that can be included without editing (e.g., photos, imaging data, etc.) can be submitted as raster files (.jpg, .png or .tif). Other figures (e.g., graph/bar charts or complex illustrations) should ideally be provided as vector files (.ai, .eps or .svg) if possible, otherwise as a .jpg, .pdf or .tif. Tables/boxes should be provided as e.g., Microsoft Word or Microsoft Excel files, and must be editable. If you are uncertain whether the format of your files is appropriate, please check with the Journal Editor.
- **Resolution:** this should be as high as possible; ideally 300 dpi for a .jpg. Images that are blurry or illegible in any way will not be accepted.
- **Fonts:** please use HELVETICA 8pt

- **Abbreviations:** All abbreviations used within Figures/tables/boxes should be defined in the legend (even if previously defined in the body of the manuscript).
- **Photomicrograph:** Please ensure that **scale bars** are included in figures where appropriate (e.g., photomicrographs). Symbols, arrows or letters used in photomicrographs should contrast with the background. Please explain internal scale and identify the method of staining in photomicrographs.
- **Editing of figures:** Future Science Group journals apply the Council of Science Editors recommendations for digital images, specifically:
  - No specific feature within an image may be enhanced, obscured, moved, removed, or introduced
  - Adjustments of brightness, contrast, or color balance are acceptable if they are applied to the whole image and as long as they do not obscure, eliminate, or misrepresent any information present in the original
  - The grouping of images from different parts of the same gel, or from different gels, fields, or exposures must be made explicit by the arrangement of the figure (e.g., dividing lines) and in the text of the figure legend
  - If the original data cannot be produced by an author when asked to provide it, acceptance of the manuscript may be revoked

Future Medicine is able to offer a number of design services to authors, from polishing an existing figure to creating one from scratch (subject to fees). If you would be interested in learning more about this service, please contact Rachel Jenkins, Head of Publishing Solutions.

Chemical structures

If possible, please submit structures drawn in ISISDraw or ChemDraw format. However, chemical structures can be redrawn in-house. Please use the following conventions:

- Always indicate stereochemistry where necessary – use the wedge and hash bond convention for chiral centers and mark cis/trans bonds as such.
- Draw small peptides (up to five amino acids) in full; use amino acid abbreviations (Gly, Val, Leu, etc.) for larger peptides.
- Refer to each structure with a number in the text; submit a separate file (e.g., not pasted throughout the text) containing these numbered structures in the original chemical drawing package that you used.

Original, reproduced and adapted material

Definitions:

Original content:

- Prepared for the current manuscript, not previously published, and
- Based on the authors' original data, or where based on other's data (i.e., an original table summarising the findings of a collection of previously published studies), presented in an original format and the sources of the data clearly cited

Reproduced content:

- Content previously published in an Open Access format, or
- Content previously published by an STM publisher, or
- Content previously published by a non-STM publisher

Adapted content:

- Content adapted from a previously published figure/table

Permissions for reproduced or adapted material within your article

Original content clearly does not require any permission to publish, however authors should ensure they cite any data sources used to create the content.

If a figure, table or box has been published previously (even if you were the author), acknowledge the original source and submit written permission from the copyright holder to reproduce the material where necessary.

When is permission required?

Content previously published in an Open Access format:

It is generally acceptable to reproduce material published in an Open Access format; however, authors should be sure to check the license under which the content was published to check this is the case, and to ensure any citation requirements are met. Further information on the types of open access license can be found here: https://creativecommons.org/.

Content previously published by an STM publisher:

Future Medicine is a signatory to the STM Permissions Guidelines produced by the International Association of Scientific, Medical and Technical Publishers (http://www.stm-assoc.org/). Permission is, or in the case of an express permission requirement should be, granted free of charge by signatory organizations, with respect to a particular journal article or book being prepared for publication, to:

- Use up to three figures (including tables) from a journal article or book chapter, but:
  - not more than five figures from a whole book or journal issue/edition;
  - not more than six figures from an annual journal volume; and
  - not more than three figures from works published by a single publisher for an article, and not more than three figures from works published by a single publisher for a book chapter (and in total not more than thirty figures from a single publisher for re-publication in a book, including a multi-volume book with different authors per chapter).
- Use single text extracts of less than 400 words from a journal article or book chapter, but not more than a total of 800 words from a whole book or journal issue/edition.

Permission to go beyond such limits may be sought although in such instances the permission grant may require permission fees. Important – although permission may be granted without charge, authors must ensure that appropriate permission has nevertheless been obtained. Co-signatories of the permissions agreement can be found on the following website: https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

Content previously published by a non-STM publisher:

As the author of your manuscript, you are responsible for obtaining permissions to use material owned by others. Please note, this is generally the publisher of the work, so even if you are the author of the previous publication, you should check the specific journal's re-use policies. Since the permission-seeking process can be remarkably time-consuming, it is wise to begin seeking permission as soon as possible. It is often possible to apply to permission to re-use content directly via an article's webpage; if you have any difficulty locating the correct place to apply for permission, please advise the Journal Editor as soon as possible.

Adapted content:

For copyright purposes, adaptation is defined as addition of substantial new information that was not previously included in the original (so for example, adding an extra line to a previously published table would likely not be considered sufficient adaptation to consider the new table 'original'). If in doubt, it is always best to err on the side of caution and obtain permission for re-use, on the same basis as described above for the various types of previously published content.

What should I submit to the journal to show permission has been obtained?

Please send us copies of letters or forms granting you permission for the use of copyrighted material so that we can see that any special requirements with regard to wording and placement of credits are fulfilled. Keep the originals for your files. If payment is required for use of the figure, this should be covered by the author.

Supplementary materials

All Future Science Group journals conduct double-anonymized peer review. Authors should ensure any supplementary materials files are anonymized (including the file names) to allow anonymity of the authors.

File size: **Total files uploaded to the journal site cannot exceed 100MB. If you have files exceeding this limit, or have any difficulty uploading your files, please contact the Journal Editor for assistance.**

Figure, tables and boxes larger than one A4 page will be included as online-only supplementary information. At the Editor's discretion data or experimental details can also be included.

Supplementary materials on Figshare

Figshare is a repository where users can make all of their research outputs, in any file format, available in a citable, shareable and discoverable manner. Future Science Group partners with Figshare to make all our supplementary files freely available and citable in their own right. Supplementary files feature both within the article on the journal site and on the Future Science Group Figshare porta