

https://dspace.org/

Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations

Malinga, Thandeka Vuyiswa Bongiwe

2024

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/43198

Malinga, Thandeka Vuyiswa Bongiwe . (2024). Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations [Master's dissertation, University of the Witwatersrand, Johannesburg]. WireDSpace.

Downloaded from a DSpace repository. https://dspace.org/

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations

Thandeka Vuyiswa Bongiwe Malinga

Student number: 2152565

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

February 2024

Supervisor: Dr David Twesigomwe

Co-supervisor: Dr Houcemeddine Othman

NATIONAL HEALTH LABORATORY SERVICE

SDIMB SYDNEY BRENNER INSTITUTE FOR MOLECULAR BIOSCIENCE

Declaration

I, Thandeka Vuyiswa Bongiwe Malinga, 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.

Signature: 

Date: 26/02/2024

Presentations arising from this research project

1. Malinga T., Othman H., Twesigomwe D. Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations. American Society of Human Genetics (Washington DC) Poster presentation. 1-5 November 2023
2. Malinga T., Othman H., Twesigomwe D. Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations. Molecular Bioscience Research Thrust Annual Research Symposium (Wits): Poster abstract accepted. 7 December 2023

Abstract

Tuberculosis (TB) is a major health burden in Africa. Although TB is treatable, anti-TB drugs are associated with adverse drug reactions (ADRs) which are partly attributed to pharmacogenetic variation. The distribution of star alleles (haplotypes) influencing anti-TB drug metabolism, is unknown in many African populations. This presents challenges in implementing genotype-guided therapy in Africa to decrease the occurrence of ADRs and enhance the efficacy of anti-TB drugs. Therefore, this study aimed to characterise the distribution of star alleles in genes that are involved in anti-TB drug metabolism (mainly isoniazid), namely *CYP2E1*, *NAT1*, *NAT2*, *GSTM1* and *GSTT1*, across diverse African populations. We used 794 high-depth whole genome sequence datasets representative of eight Sub-Saharan African (SSA) population groups. Data sources included the 1000 Genomes Project and H3Africa AWi-Gen. *CYP2E1*, *NAT1*, *NAT2*, *GSTM1* and *GSTT1* star alleles were called from the WGS data using StellarPGx. Subsequently, novel star alleles were analysed, and their allele defining variants were annotated using the Ensembl Variant Effect Predictor. We present the distribution of both common and rare star alleles influencing anti-TB drug metabolism across various SSA populations, in comparison to other global populations. Various key star alleles were identified in the SSA study populations at relatively high frequencies including *NAT1*10*, *GSTT1*0* (>50%), *GSTM1*0* (49%), and *NAT2*5B* (21%). Additionally, we predicted varying phenotypic proportions for NAT1 and NAT2 (acetylation) and the GST enzymes (detoxification activity) between SSA and other global populations. Fifty potentially novel haplotypes were identified computationally across the five genes. This study provides insight into the distribution of star alleles in genes relevant to isoniazid metabolism across various African populations. The high number of potentially novel star alleles exemplifies the need for pharmacogenomics studies in the African context. Overall, our analysis provides a foundation for implementing pharmacogenetic testing in Africa to reduce the risk of ADRs related to TB treatment.

Acknowledgement

I would like to thank my supervisors, David Twesigomwe and Houcemeddine Othman. David, it goes without saying how much gratitude I have in the manner that you have guided me this year, particularly with this research report. Thank you for the many conversations, always being willing to help, and most importantly introducing me to the once daunting world of bioinformatics. Houcem, thank you for the many opportunities you have provided including being able to join a research project that can make great strides in anti-TB treatment optimisation, the MSc research fellowship, and the openness in allowing me to travel to present my work. I have truly learnt a lot, so thank you.

I would also like to extend my thanks to Wits University for the Postgraduate Merit Award, which aided in funding my postgraduate studies.

Thank you to everyone at the Division of Human Genetics, including staff and the MSc (Med) Genomic Medicine 2023 class. This year has been filled with learning what is new in human genetics, this knowledge has put me in great stead in advancing my career in the field of human genetics.

Thank you to the scientific community at the Sydney Brenner Institute of Molecular Bioscience, particularly the AGORA-TM team, for their help during this year.

Lastly, to my dear parents, thank you for all that you have done for me. I am immensely grateful for the encouragement and advice that you both are always so willing to offer, especially during this eventful academic year. To my brothers, thank you for supporting me in your own unique ways. Thank you to God for blessing me with such a supportive network of friends, classmates, supervisors, and family.

Table of contents

Introduction 1

Methods 5

- Study design 5
- Ethics 5
- Data sources 5
- Star allele bioinformatics analysis 7
- Phenotype prediction 8
- Variant annotation and functional prediction 9
- Statistical analysis 9

Results 9

- StellarPGx Benchmarking Analysis 9
- Summary of variants 10
- CYP2E1 12
  - Distribution of *CYP2E1* non-synonymous variants across SSA and various global populations 12
  - *CYP2E1* haplotypes characterised using trio data 16
  - *CYP2E1* structural variants among the SSA populations 18
- NAT1 and NAT2 19
- Star allele frequencies 19
  - Potentially novel *NAT1* and *NAT2* haplotypes and diplotypes 24
  - NAT1 and NAT2 phenotype predictions 28
- GSTM1 and GSTT1 30
  - Star allele frequencies 30
  - Potentially novel *GSTM1* and *GSTT1* star alleles 33
  - GSTM1 and GSTT1 predicted phenotypes 36

Discussion 38

Future directions 42

Conclusion 43

References 44

Appendices

A. Approved research protocol
B. Ethics clearance certificate
C. Turn-it-in report
D. Author guidelines for *Frontiers in Pharmacology*

List of Figures

Figure 1: Isoniazid (INH) metabolism in the liver 2
Figure 2: Key star alleles and the corresponding allele defining variants in five anti-TB pharmacogenes: *NAT2* (a), *NAT1* (b), *CYP2E1* (c), *GSTM1* (d), and *GSTT1* (e)..... 4
Figure 3: Summary of the bioinformatics analysis workflow for the study 7
Figure 4: StellarPGx performance on calling star alleles from 70 WGS samples (from the Polaris HiSeqX pharmacogenomics cohort)..... 10
Figure 5: A summary of the variation observed across *CYP2E1*, *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* in this study 11
Figure 6: Characterisation of *CYP2E1* (NG_008383.1) haplotypes from trio data. 17
Figure 7: *CYP2E1* copy number variants identified in this study..... 18
Figure 8: Characterisation of a potentially novel *NAT1* haplotype using inheritance information 25
Figure 9: Distribution of NAT1 and NAT2 phenotype predictions across Sub-Saharan Africa (SSA) compared to other global populations. 29
Figure 10: Read depth of *GSTM1* structural variants..... 31
Figure 11: Distribution of GSTM1 (a) and GSTT1 (b) phenotype predictions across SSA compared to other global populations 37

List of Tables

Table 1: WGS data of various African populations/countries analysed with their respective data sources 6
Table 2: Frequencies of *CYP2E1* functionally-relevant variants in SSA compared with various global populations 13
Table 3: Distribution of *CYP2E1* variants across SSA populations in this study 15
Table 4: The distribution of *NAT1* and *NAT2* star alleles in SSA compared with various global populations 21
Table 5: The distribution of *NAT1* and *NAT2* star alleles variants across SSA populations in this study..... 24
Table 6: Possible novel *NAT1* and *NAT2* haplotypes and unresolved diplotypes in the SSA populations 26
Table 7: The distribution of *GSTM1* and *GSTT1* star alleles in SSA compared with various global populations 32
Table 8: The distribution of *GSTM1* and *GSTT1* star alleles variants across SSA populations in this study..... 33

Table 9: Possible novel *GSTM1* and *GSTT1* haplotypes in the SSA populations 34
Supplementary Table 1: Diplotype calls made by StellarPGx and Stargazer in comparison to reference materials from the CDC's Genetic Testing Reference Materials Coordination Program (GeT-RM) 49
Supplementary Table 2: *In-silico* predictions made for prioritised allele defining variants for known and possible novel star alleles 52
Supplementary Table 3: Examples of trio data used to infer core *CYP2E1* (NG_008383.1) variants defining novel haplotypes 56
Supplementary Table 4: The distribution of *NAT2* star alleles with unknown acetylation effects in SSA compared with various global populations 69
Supplementary Table 5: The distribution of *NAT2* star alleles with unknown acetylation effects across SSA populations in this study..... 72
Supplementary Table 6: Examples of trio data used to infer core *NAT1* (NG_012245.2) and *NAT2* (NG_012246.1) variants defining novel haplotypes 73
Supplementary Table 7: The distribution of *NAT1* diplotypes and NAT1 predicted phenotypes in SSA compared with various global populations 76
Supplementary Table 8: The distribution of *NAT2* diplotypes and NAT2 predicted phenotypes in SSA compared with various global populations 78
Supplementary Table 9: The distribution of *GSTM1* diplotypes and the associated GSTM1 enzyme activity phenotypes in SSA compared with various global populations 84
Supplementary Table 10: The distribution of *GSTT1* diplotypes and associated GSTT1 enzyme activity phenotypes in SSA compared with various global populations 84

List of abbreviations

ADRs	Adverse drug reactions
AF	Allele frequency
AMR	Admixed American
AT-DILI	Anti-TB drug-induced liver injury
BFA	Burkina Faso
CNVs	Copy number variants
CPIC	Clinical Pharmacogenetics Implementation Consortium
CYP2E1	Cytochrome P450 2E1
EAS	East Asian
ESN	Esan in Nigeria
EUR	European
GeT-RM	Genetic Testing Reference Materials Coordination program
GHA	Ghana
GSTM1	Glutathione S-transferase Mu1
GSTT1	Glutathione S-transferase Theta-1
GWD	Gambian in Western Division, Mandinka
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
INH	Isoniazid
LWK	Luhya in Webuye, Kenya
MSL	Mende in Sierra Leone
NAT1	N-Acetyltransferase 1
NAT2	N-Acetyltransferase 2
PAS	p-Aminosalicyclic acid
SNVs	Single nucleotide variants
SAS	South Asian
SEB	Southeastern Bantu (SEB) in South Africa
SSA	Sub-Saharan African
SV	Structural variants
TB	Tuberculosis
WGS	Whole genome sequence
YRI	Yoruba in Ibadan

Research report in the format of a “submissible” paper

Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations

Thandeka V.B. Malinga [1,2], Houcemeddine Othman [2], David Twesigomwe [1,2]

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

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

Manuscript to be submitted to *Frontiers in Pharmacology*

Keywords: Africa, anti-tuberculosis drugs, pharmacogenomics, star alleles, drug-induced liver injury, precision medicine

Abstract for journal submission

Background: Tuberculosis (TB) is a major health burden in Africa. Although TB is treatable, anti-TB drugs are associated with adverse drug reactions (ADRs) which are partly attributed to pharmacogenetic variation. The distribution of star alleles (haplotypes) influencing anti-TB drug metabolism, is unknown in many African populations. This presents challenges in implementing genotype-guided therapy in Africa to decrease the occurrence of ADRs and enhance the efficacy of anti-TB drugs. **Methods:** In this study, StellarPGx was used to call star alleles in five anti-TB pharmacogenes, namely *CYP2E1, NAT1, NAT2, GSTM1* and *GSTT1,* from 794 high-depth whole genome sequence datasets, representative of eight Sub-Saharan African (SSA) ethnolinguistic groups. Subsequently, novel star alleles were analysed, and their allele defining variants were annotated using the Ensembl Variant Effect Predictor. **Results:** We present the distribution of both common and rare star alleles influencing anti-TB drug metabolism across various SSA populations, in comparison to other global populations. Various key star alleles were identified in the SSA study populations at relatively high frequencies including *NAT1*10*, *GSTT1*0* (>50%), *GSTM1*0* (49%), and *NAT2*5B* (21%). Additionally, we predicted varying phenotypic proportions for NAT1 and NAT2 (acetylation) and the GST enzymes (detoxification activity) between SSA and other global populations. Fifty potentially novel haplotypes were identified computationally across the five genes. **Conclusion:** This study provides insight into the distribution of star alleles relevant to anti-TB drug metabolism across various African populations. The high number of potentially novel star alleles exemplifies the need for pharmacogenomics studies in the African context. Overall, our analysis provides a foundation for implementing pharmacogenetic testing in Africa to reduce the risk of ADRs related to TB treatment.

Introduction

Tuberculosis (TB) is a major public health concern in many developing countries worldwide. Over 15 African countries, including South Africa, Nigeria, and Kenya, are characterised by high TB incidence and mortality (WHO, 2022). This is mainly due to various socioeconomic factors and comorbidities involving HIV/AIDS and other prevalent communicable diseases in Africa (Boyce et al., 2019). Historically, TB treatment has been defined by administering four first-line anti-TB drugs: isoniazid (INH), rifampin, pyrazinamide, and ethambutol (WHO, 2022). Most studies on combating TB have been focused on addressing the problem of drug-resistant TB, with newer treatments such as bedaquiline becoming increasingly beneficial and popular in this regard (Haas et al., 2022; Toure et al., 2016). However, another existing challenge is the considerable variability in interindividual responses to these antimycobacterial drugs, often leading to poor treatment outcomes and undermining TB eradication efforts. On one hand, some individuals may experience either no response or a subtherapeutic response to standard doses of current anti-TB drugs. On the other hand, some patients develop severe adverse drug reactions (ADRs) such as hepatotoxicity, thus prompting interruptions to treatment (Costa et al., 2012; Zenebe et al., 2016), and an increased risk of drug-resistant strains emerging.

Anti-TB drug-induced liver injury (AT-DILI) is a serious ADR to anti-TB drugs resulting from the damaging effects of toxic metabolites (Chen et al., 2015). The global incidence of AT-DILI varies from 0.1-27.7% (Chen et al., 2015), and a similar variance has been observed in different African countries (Abera et al., 2016; Kesenogile et al., 2021). AT-DILI, and overall variability in anti-TB drug response, have been attributed to both genetic and non-genetic factors, including age, comorbidities, drug-drug interactions, and lifestyle habits (Cavaco et al., 2022). In particular, genetic variation in drug-metabolism-related pharmacogenes, namely, *CYP2E1, NAT1, NAT2, GSTM1,* and *GSTT1*, have been associated with an increased AT-DILI risk (Levano et al., 2021). Therefore genotype-guided treatment for individuals with TB may aid in reducing the risk of ADRs to anti-TB drugs and increase treatment efficacy (Azuma et al., 2013) in high-TB-burden African countries.

CYP2E1 (phase I metabolising enzyme), NAT2, GSTM1, and GSTT1 (phase II metabolising enzymes) are key enzymes in the metabolism of first-line anti-TB drugs (reviewed by Sanni et al., 2022). For example, the main metabolic pathway for INH entails acetylation by NAT2 (Figure 1). The encoding *NAT2* gene is highly polymorphic, resulting in diplotypes that correspond to three different NAT2 acetylation effects: rapid, intermediate, and slow acetylation. In the secondary INH pathway, CYP2E1 metabolises INH to toxic metabolites that require clearance mainly via GSTM1 and GSTT1. Therefore, the occurrence of increased

function of *CYP2E1* alleles and/or deletion(s) of *GSTM1* and *GSTT1* may contribute to the risk of AT-DILI (Sanni et al., 2022). An additional phase II metabolising enzyme is NAT1 which acetylates p-aminosalicyclic acid (PAS), a second-line anti-TB drug used in individuals with TB that are resistant to first-line anti-TB drugs (Sy et al., 2015).

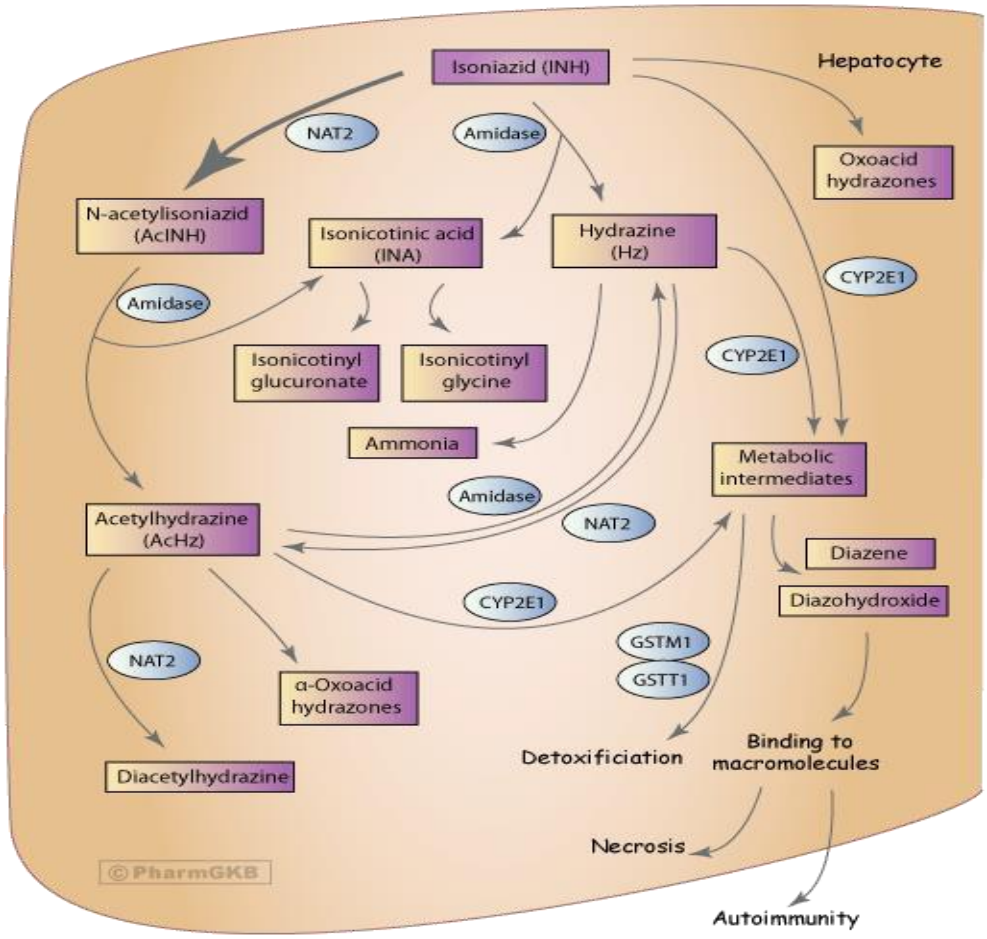


Figure 1: Isoniazid (INH) metabolism in the liver. INH can undergo acetylation, hydrolysis of an amide group or oxidation by NAT2, amidase or CYP2E1 enzymes, respectively. The INH metabolites are detoxified by GSTM1 and GSTT1. Image credit: PharmGKB (https://www.pharmgkb.org/pathway/PA166151813) shared under Creative Commons BY-SA 4.0 license.

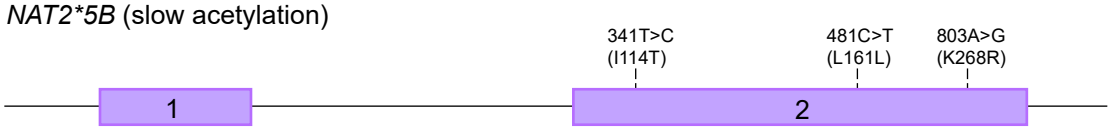
Genetic variation in *CYP2E1*, *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* consists of various combinations (haplotypes) of single nucleotide variants (SNVs), small insertions and deletions (indels), and/or structural variants (SVs). As with most major pharmacogenes, the star (*) allele nomenclature is used to represent these haplotypes. With exception to most reference haplotypes, star alleles are typically defined by missense variants, nonsense variants, frameshift variants, splicing defects, regulatory variants, and/or copy number variants (CNVs) (Figure 2).

a *NAT2* – 8p22 (NM_000015.3)

*NAT2*12A* (rapid acetylation)

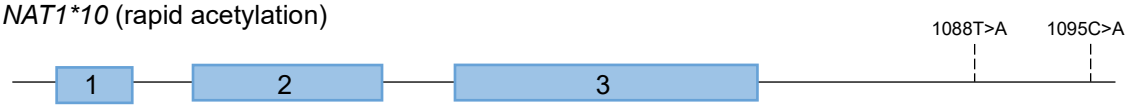


*NAT2*5B* (slow acetylation)

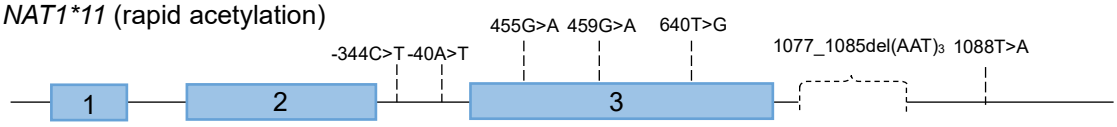


b *NAT1* – 8p22 (NM_000662.8)

*NAT1*10* (rapid acetylation)



*NAT1*11* (rapid acetylation)



c *CYP2E1* – 10q26.3 (NM_000773.4)

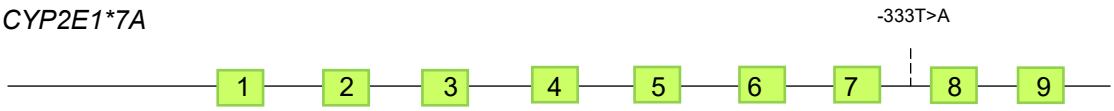
*CYP2E1*1A* (reference)



*CYP2E1*1C*



*CYP2E1*7A*

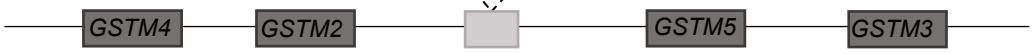


d *GSTM1* – 1p13.3 (NM_000561.4)

*GSTM1*A* (reference)



*GSTM1*0* (gene deletion)



e *GSTT1* – 22q11.23 (NM_000853.4)

*GSTT1*A* (reference star allele)

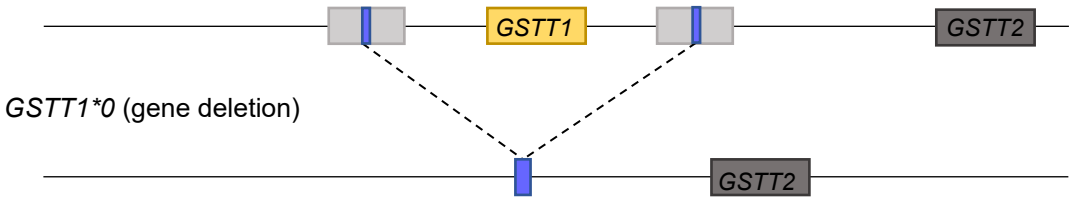


Figure 2: Key star alleles and the corresponding allele defining variants in five anti-TB pharmacogenes: *NAT2* (a), *NAT1* (b), *CYP2E1* (c), *GSTM1* (d), and *GSTT1* (e).

A number of these star alleles have been associated with an increased risk of AT-DILI in various populations (Jaramillo-Valverde et al., 2022). For example, individuals homozygous for *NAT2*5B* (slow acetylation) (El-Jaick et al., 2022), *GSTM1*0,* and *GSTT1*0* (gene deletion) (Chbili et al., 2022) are associated with AT-DILI in patients treated with first-line anti-TB drugs.

Based on the ever-expanding pharmacogenetics body of knowledge over the last three decades, consortia such as the Clinical Pharmacogenetics Implementation Consortium (CPIC) (https://cpicpgx.org/) have established genotype-guided drug dosing recommendations with a view toward reducing unfavourable treatment responses. However, no consensus CPIC genotype-guided dosing guidelines have been developed in relation to TB treatment thus far, despite the evidence that several star alleles are known to be associated with an increased risk of AT-DILI. The absence of standardised genotype-guided dosing guidelines hinders personalised treatment strategies across Africa and other genetically diverse regions where TB and anti-TB-related ADRs are highly prevalent. A major challenge in establishing consensus genotype-guided dosing guidelines and the appropriate pharmacogenetic testing strategies, is the deficit in information regarding the distribution of star alleles impacting the function of anti-TB drug metabolising enzymes, especially across Africa. This is partly due to the underrepresentation of various African ethnolinguistic groups in genomics studies to date, particularly in pharmacogenomics studies. That being said, due to recent advances in next generation sequencing (NGS) technologies that have rendered genome sequencing cheaper and more accessible, there is an increasing availability of whole genome sequence (WGS) datasets in Africa. Furthermore, the Human, Heredity, and Health in Africa (H3Africa) consortium, and other Africa-based collaborative projects, have enabled data generation efforts in previously understudied populations. This provides an opportunity to address the gaps in knowledge on star allele distributions pertinent to anti-TB drug response and may further expand to other drug-gene pairs.

Therefore, this study aims to leverage recently generated African WGS datasets to characterise the distribution of star alleles in genes that are involved in anti-TB drug metabolism (mainly isoniazid), namely *CYP2E1, NAT1, NAT2, GSTM1* and *GSTT1*, across diverse African populations. This study adds to the growing efforts to enhance precision medicine in Africa in the context of TB treatment. Understanding the landscape of common, rare and/or novel star alleles in these genes could inform optimisation of anti-TB drug dosage regimens for patients at risk of AT-DILI or treatment non-response in Africa and the African-diaspora.

Methods

Study design

This was a comparative study entailing secondary analysis of WGS data to characterise the star allele distribution in pharmacogenes involved in anti-TB drug metabolism across various Sub-Saharan African (SSA) populations, and in comparison with global populations. Given that the study datasets were all short-read high coverage (≥30x) WGS, we included a pilot study to assess the accuracy of using StellarPGx (Twesigomwe et al., 2021) for calling *CYP2E1, NAT1, NAT2, GSTM1* and *GSTT1* star alleles from this type of input data before progressing to the full-scale analysis. Given the lack of current comprehensive star allele definitions for *CYP2E1*, in the main pharmacogenomic analysis across SSA populations we analysed variation in this gene independent of the star allele annotations.

Ethics

Ethical clearance for this study was granted by the University of the Witwatersrand (Wits) Human Research Ethics Committee (Medical) under protocol number: M230687 and as a substudy of a parent study titled “Approaches of GenOmics for Response to drugs in Africa-TB-Malaria” (AGORA-TM) (M220738). Permission to use non-publicly available WGS datasets was obtained from the respective data gatekeepers. Ethics approval was not required for the publicly available datasets.

Data sources

For the StellarPGx benchmarking study, we used high coverage (30x) WGS data of 70 participants within the Polaris HiSeqX pharmacogenomics cohort (https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort). These include participants with African ancestry (n=19), East Asian ancestry (n=19), European ancestry (n=24), and admixed American participants (n=7) (Pratt et al., 2016). We used *CYP2E1*, *NAT1, NAT2, GSTM1* and

GSTT1 consensus star allele calls confirmed by the CDC's Genetic Testing Reference Materials Coordination (GeT-RM) Program, in previous pharmacogenetics analysis (Pratt et al., 2016), as gold-standard calls. In addition, StellarPGx diplotype calls were compared with Stargazer calls (Lee et al., 2019) made using the same data.

For the main research question, high coverage WGS data (~30x) of 794 African individuals were analysed from five datasets (Table 1). These datasets consisted of SSA populations, and the participants were predominantly from Niger-Congo language family ethnolinguistic groups.

Table 1: WGS data of various African populations/countries analysed with their respective data sources

African countries/populations	**Primary project**	**n** (794)
Public genomic datasets		
Gambian in Western Division, Mandinka (GWD)	1000 Genome Project	113
Yoruba in Ibadan (YRI)	1000 Genome Project	108
Esan in Nigeria (ESN)	1000 Genome Project	99
Luhya in Webuye, Kenya (LWK)	1000 Genome Project	99
Mende in Sierra Leone (MSL)	1000 Genome Project	85
Various African ethnolinguistic groups	SGDP	31
Genomic datasets from Africa-based research projects		
Southeastern Bantu (SEB) in South Africa	AWI-Gen	100
Burkina Faso (BFA)	AWI-Gen	33
Ghana (GHA)	AWI-Gen	26
South Africa	CBRL (NICD, Wits)	85
South Africa	SAHGP	15

SGDP - Simons Genome Diversity Project, AWI-Gen - Africa Wits-INDEPTH partnership for Genomics, SAHGP - South African Human Genome Project, CBRL (NICD, WITS) - Cell Biology Research Lab, National Institute of Communicable Diseases, University of the Witwatersrand; n – population sample size

The 1000 Genomes Project (1000G) (Byrska-Bishop et al., 2022) public dataset consisted of 504 genomes (30x) representative of five different SSA populations (Table 1). Additional 1000G populations were included in this analysis to compare star allele frequencies in the SSA populations with other global populations. These global populations comprised of African Caribbean participants in Barbados (ACB; n=96), participants with African Ancestry in Southwest USA (ASW; n=61), admixed American participants (AMR; n=347), East Asian participants (EAS; n=504), European participants (EUR; n=503), and South Asian participants (SAS; n=489).

The public dataset from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016) included 31 genomes (43x) of participants from various SSA countries, including the Congo,

Namibia, Kenya, Senegal, Algeria, Nigeria, Sudan, and South Africa. Three samples had ambiguous country designations and were documented as Botswana/Namibia by the SGDP.

The Africa Wits-INDEPTH partnership for Genomics (AWI-Gen) datasets comprised of 59 west African WGS samples from Burkina Faso (BFA; n=33), and Ghanaian (GHA; n=26) participants sequenced as part of the H3Africa-Baylor project (Ramsay et al., 2016), and 100 genomes (40x) from southeastern Bantu-speaking participants (SEB). The WGS data (~30x) obtained from the Southern African Human Genome Programme (SAHGP) comprised of genomes from 15 South African individuals further delineated into two Bantu-speaking populations: Sotho-Tswana (n=8) and Xhosa (n=7) (Choudhury et al., 2017). Finally, 85 South African genomes (>30x) from the Cell Biology Research Lab affiliated with the National Institute of Communicable Diseases (NICD)/ Wits were analysed.

Star allele bioinformatics analysis

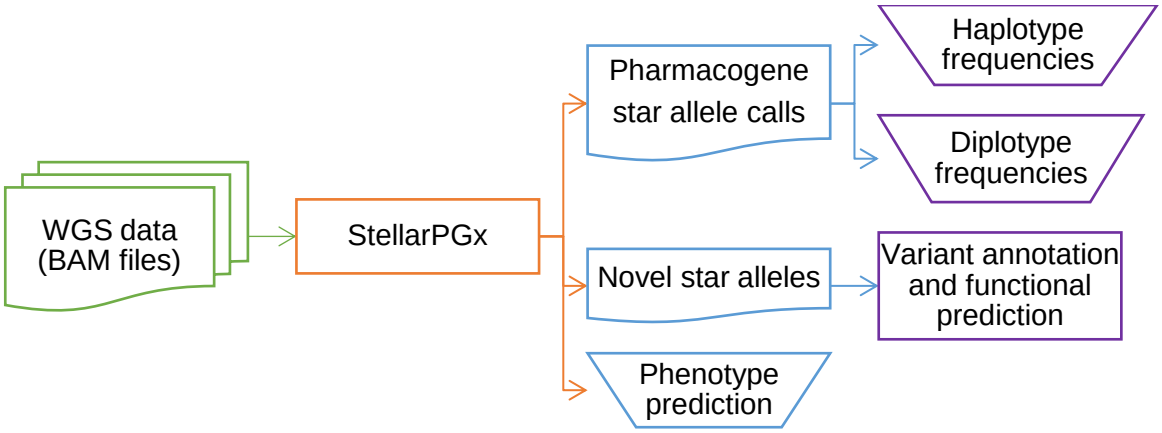


Figure 3: Summary of the bioinformatics analysis workflow for the study.
BAM, Binary Alignment Map.

Star alleles in *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* were called using StellarPGx (Twesigomwe et al., 2021), a Nextflow (Tommaso et al., 2017) environment, by supplying WGS Binary Alignment Map (BAM) files aligned to the GRCh38 reference genome as input (Figure 3). For *CYP2E1*, only structural variants were detected using StellarPGx due to aforementioned star allele definition challenges. Briefly, in StellarPGx v1.2.6, variants are called via GraphTyper, which performs read realignment to a variant-aware reference genome (i.e. pangenome) to improve the accuracy of subsequent variant calling steps. The GraphTyper output consists of variant genotype calls and allele depths stored in variant call format (VCF) files.

After variant calling, StellarPGx assigns candidate star alleles by comparing the combinations of allele defining SNVs in the VCFs to known diplotypes in the StellarPGx database. In some cases, the core SNV combination from the VCF matches more than one diplotype in the database. This is resolved within StellarPGx by considering sub-allele variants (non-function-altering variants in a haplotype) to determine the appropriate candidate diplotype. Typically, the StellarPGx database consists of precomputed diplotype combinations based on star allele definitions curated by PharmVar. However, since *CYP2E1, NAT1*, *NAT2*, *GSTM1*, and *GSTT1* had not yet been transitioned to the PharmVar database at the time of this study, we used *NAT1* and *NAT2* star allele definitions from the Database of arylamine *N*-acetyltransferases (NATs) (http://nat.mbg.duth.gr/), *GSTM1*, and *GSTT1* allele designations from previous literature (Townsend & Tew, 2003), while previously published *CYP2E1* star allele definitions were referred to on a case-by-case basis.

Star alleles consisting of CNVs were called by comparing the depth of coverage of the target gene, obtained from SAMtools *depth*, to the read depth of two control loci, vitamin D receptor (*VDR)* and epidermal growth factor receptor (*EGFR*), to elucidate the copy number (CN) changes in the key gene regions. These control loci (CN=2) were chosen because they are relatively large genes with fundamental biological functions and are therefore less likely to be entirely affected by CNVs. This was also supported by data from the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). This information was then combined with SNV allele depth information in the VCF files to ascertain the nature of the duplications, deletions, and other structural rearrangements.

For samples with novel star alleles or sub-alleles, StellarPGx provides a flag in the output, and in addition it displays the core variants, and the possible background diplotype. This narrows down the number of samples that require additional manual inspection (e.g. visualisation on the Integrative Genomics Viewer) of novel star allele defining variants.

Phenotype prediction

The phenotype prediction was based on combining the associated activity of the haplotypes that form an individual's diplotype for a particular gene. For *NAT2*, we adopted the diplotype-phenotype translation used previously by Sabbagh et al. (2011) and other studies, whereby three phenotype categories are possible i.e. NAT2 slow acetylator (presence of two star alleles associated with slow acetylation e.g. *NAT2*5B*), NAT2 intermediate acetylator (presence of one slow acetylation star allele and one fast acetylation star allele), and NAT2 fast acetylator (presence of two star alleles associated with rapid acetylation e.g. *NAT2*4*). A similar

approach was adapted to infer NAT1, GSTM1 and GSTT1 phenotypes, respectively. However, potential CYP2E1 phenotypes were not predicted in this study due to challenges with existing star allele definitions (see Results and Discussion).

Variant annotation and functional prediction

Alleles defining novel haplotypes were annotated using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016). The *in-silico* VEP plugins used for functional predictions of allele defining variants included CADD (Rentzsch et al., 2019), LRT (Fujita et al., 2010), Mutation Assessor (Reva et al., 2011), PolyPhen-2 (Adzhubei et al., 2013), PROVEAN (Choi & Chan, 2015), SIFT (Ng & Henikoff, 2003), and VEST4 (Carter et al., 2013). Prediction scores were based on pharmacogenetic-optimised scores (Zhou et al., 2019) for more accuracy. A variant was considered deleterious if >4 *in-silico* prediction tools predicted it to be deleterious. Additionally, DynaMut2 (Rodrigues et al., 2021) was used to predict their potential impact on the stability of the respective protein structures.

Statistical analysis

Hardy-Weinberg Equilibrium (HWE) was determined using the genetics package in R programming (https://www.r-project.org/) to detect deviations from the expected frequencies. Fisher's exact test was used to assess the statistical significance of any observed frequency differences across the populations. The p-values were adjusted according to the BY method (Benjamini & Yekutieli, 2001) and a cut-off of 0.05 was used.

Results

StellarPGx Benchmarking Analysis

For the Polaris HiSeqX dataset, Supplementary Table 1 has the full diplotype calls made by StellarPGx, Stargazer, and the corresponding consensus 'gold standard' calls from the GeT-RM study. The relative haplotype and diplotype concordance were the main criteria used to assess the performance of StellarPGx. In this context, concordance refers to the percentage of haplotype/diplotype calls made by StellarPGx, that matched with the GeT-RM consensus calls ('gold standard') (Pratt et al., 2016), or Stargazer calls (Lee et al., 2019), respectively.

The haplotype and diplotype concordance between StellarPGx GeT-RM 'gold standard' calls varied per gene, with *NAT2* having the highest concordance (>85%) while *NAT1* (60%) and *CYP2E1* (56.4%) were relatively low (Figure 4, Supplementary Table 1). A relatively similar observation was seen for the haplotype and diplotype concordance between GeT-RM and

Stargazer (Figure 4, Supplementary Table 1). In contrast, a high haplotype (>97%) and diplotype (>94%) concordance between StellarPGx and Stargazer was observed in *NAT1, NAT2, GSTM1,* and *GSTT1* (Figure 4, Supplementary Table 1). Complete concordance with Stargazer was not reached for *NAT1* and *GSTT1*, due to 10 samples with possible novel star alleles identified by StellarPGx (Supplementary Table 1) that were not flagged by Stargazer (Lee et al., 2019). Due to the above observations, StellarPGx was considered sufficient and appropriate for the subsequent main analysis in this study especially for *NAT1*, *NAT2*, *GSTM1*, and *GSTT1*. For *CYP2E1*, given the large inconsistencies observed due to lack of a well-curated star allele nomenclature, StellarPGx was used only for structural variant detection, additionally we opted to report the SNVs and frequently observed haplotypes in this gene without using any allele nomenclature system.

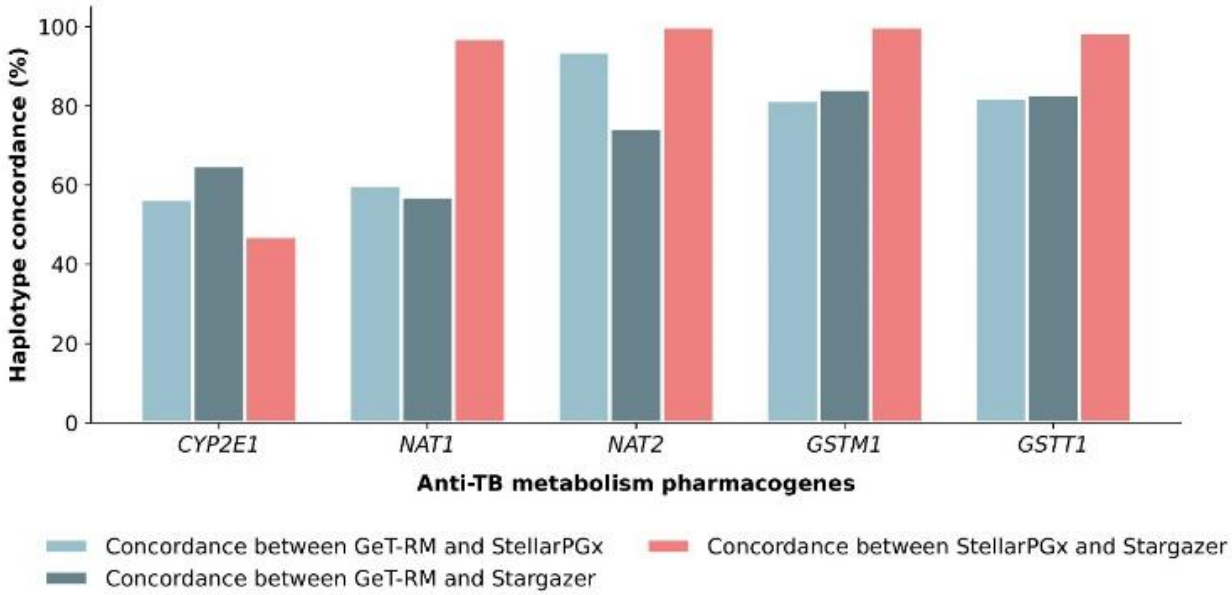


Figure 4: StellarPGx performance on calling star alleles from 70 WGS samples (from the Polaris HiSeqX pharmacogenomics cohort). 'Gold standard' star allele calls are available through the CDC's Genetic Testing Reference Materials Coordination (GeT-RM) Program.

Summary of variants

A total of 1450 SNVs were identified across the five anti-TB pharmacogenes in the SSA populations analysed in this study. The breakdown of the variant consequences is shown in Figure 5a. The SSA study population had the highest number of variants (~24 variants per kb) across the six global populations (~13-14 variants per kb) (Figure 5b). For the individual genes, *NAT1* (~35 variants per kb) had the highest number of variants followed by *CYP2E1* (~33 variants per kb) in SSA (Figure 5c).

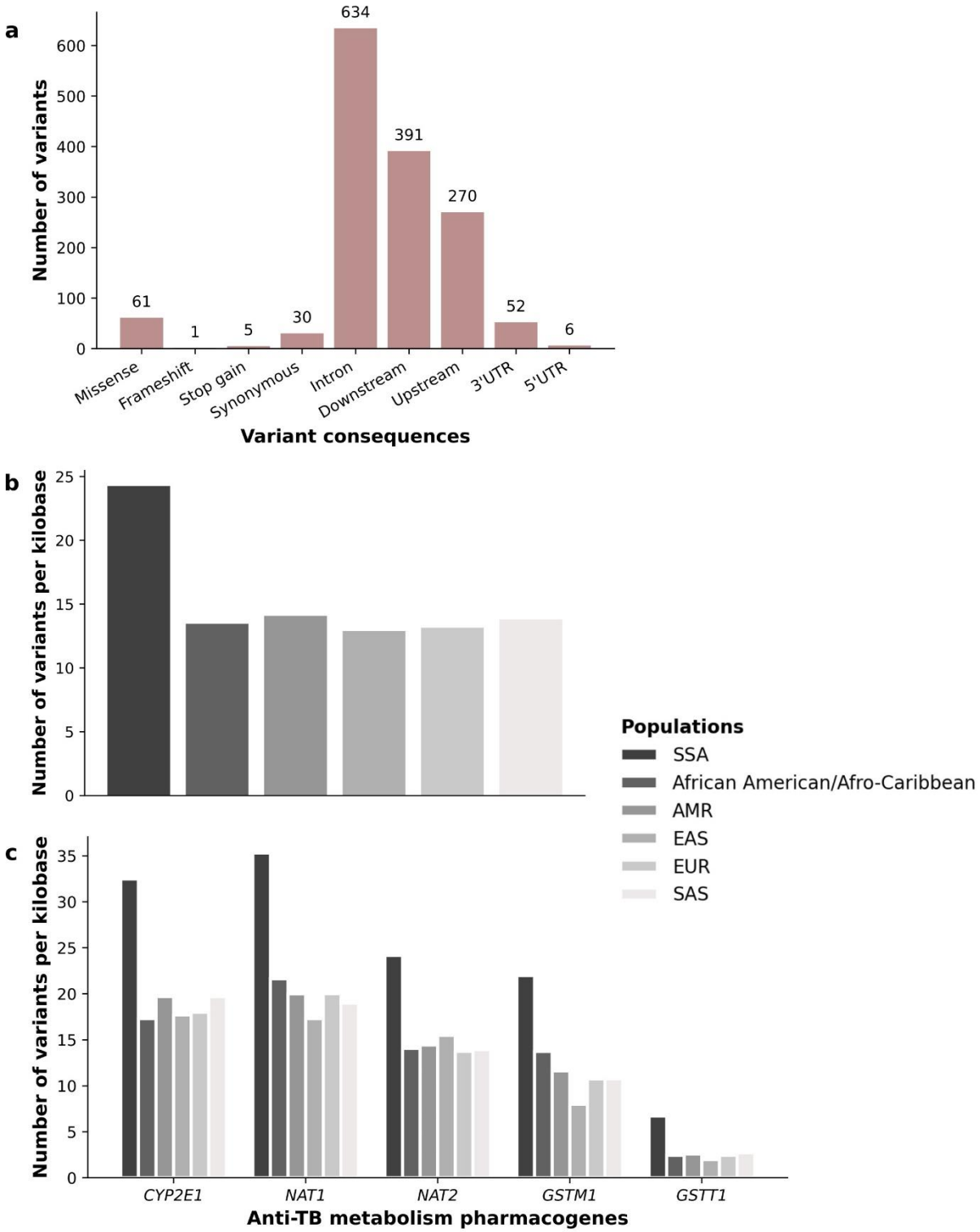


Figure 5: A summary of the variation observed across *CYP2E1*, *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* in this study. (a) Breakdown of the variant types and their count across SSA populations. (b) Comparison of the variant density in the five anti-TB metabolism pharmacogenes combined across SSA versus other global populations (c) Comparison of the variant density in each anti-TB drug metabolism pharmacogene across SSA versus other global populations analysed in this study. SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants; UTR – untranslated region

CYP2E1

Unlike other pharmacogenomic-relevant *CYP* genes, *CYP2E1* has not yet been transitioned to PharmVar. A few star alleles (**1A* to **7C*) have been published in literature, but they lack curation by an expert panel. Given this context, attempting to accurately resolve *CYP2E1* star alleles using short-read sequencing data is very challenging. Therefore, herein we highlight the distribution of *CYP2E1* non-synonymous variants, possible novel haplotypes, and structural variants, but without particular focus on the star allele nomenclature.

Distribution of *CYP2E1* non-synonymous variants across SSA and various global populations

Missense and stop-gain variants were the only two types of *CYP2E1* non-synonymous variants identified in the SSA and other global populations (Table 2). The g.5167T>C (p.Leu45Pro) variant was identified only in SSA (0.06%) and was predicted to be deleterious by majority of the *in-silico* predictions tools used in this study (Supplementary Table 2). This variant had a frequency of 0.3% in the SEB population and was absent from the remaining SSA populations groups (Table 3). The absence of this missense variant on dbSNP and gnomAD, indicates that it may possibly be a novel variant.

Of the remaining 11 missense variants identified in the various SSA populations, rs6413419 (defines **4*) was relatively common in SSA (AF=24.6%) and was statistically different from the AMR ($p=4.0x10^{-35}$), EUR ($p=1.2x10^{-55}$), and SAS ($p=1.0x10^{-72}$) populations (Table 2). Similarly, this variant was observed at high frequencies in the individual SSA populations (Table 3), with SEB (55%) having the highest frequency yet only being statistically different from the frequency observed in the LWK population (AF=13.6%; p=0.0066).

Seven non-coding SNPs define known *CYP2E1* star alleles such as **1B* (rs2070676) and **7C* (rs2070672 and rs2070673). The rs2070673 and rs2070676 variants occurred at >20% in the SSA populations. Significantly higher frequencies for both variants were observed across the global populations, particularly in the EUR populations where the rs2070673 and rs2070676 variant occurred at 81.6% and 87.2%, respectively (Table 2). The rs2031920 variant was absent in the SSA populations and occurred at low frequencies in the other global populations (<5%). The intronic variant (g.13007G>A) used to define **5A* and **6A*, was not called across the SSA and global populations within this study (Table 2).

Table 2: Frequencies of *CYP2E1* functionally-relevant variants in SSA compared with various global populations

rsID	Location (GRCh38)	Variant alleles (NG_008383.1)	Protein change	Consequence	Allele frequencies (%)					
					SSA **(n=974)**	African American/Afro-Caribbean **(n=157)**	AMR **(n=347)**	EAS **(n=504)**	EUR **(n=503)**	SAS **(n=489)**
Non-synonymous variants										
rs543066971	133527426	g.5064C>A	p.Leu11Met	Missense	0	0	0	0	0	0.1
rs367957731	133527444	g.5082C>T	p.Leu17Phe	Missense	0.06	0	0	0	0	0
rs527949682	133527527	g.5165C>A	p.Asn44Lys	Missense	0	0	0	0.10	0	0
Novel variant	133527529	g.5167T>C	p.Leu45Pro	Missense	0.06	0	0	0	0	0
rs143746211	133527550	g.5188A>G	p.Asn52Ser	Missense	0	0	0	0	0.10	0
rs35844228	133528517	g.6155G>T	p.Val72Leu	Missense	0	0	0	0.20	0	0
*rs72559710	133528530	g.6168G>A	p.Arg76His	Missense	0	0	0	0.30	0	0
rs543855005	133528614	g.6252C>T	p.Pro104Leu	Missense	0	0	0.14	0	0	0
rs60719153	133531623	g.9261C>T	p.Arg126Trp	Missense	0	0	0.14	0	0.20	0
rs200955165	133531696	g.9334A>C	p.Glu150Ala	Missense	0	0	0	0.10	0	0
rs60452492	133532153	g.9791G>A	p.Gly173Ser	Missense	0	0	0.14	0	0.10	0.92
rs56040284	133532159	g.9797G>A	p.Ala175Thr	Missense	0	0	0	0.20	0	0
rs548262477	133532160	g.9798C>T	p.Ala175Val	Missense	0	0	0.14	0	0	0
*rs6413419	133532171	g.9809G>A	p.Val179Ile	Missense	24.60	22.00	4.50	0	3.08	1.12
rs577159401	133532273	g.9911C>T	p.Pro213Ser	Missense	0	0	0	0	0	0.20
rs41299426	133532698	g.10336A>G	p.Asn219Asp	Missense	0	0	0	0.60	0	0
rs569948279	133532750	g.10388T>C	p.Ile236Thr	Missense	0	0	0	0.10	0	0
rs540609626	133532828	g.10466C>T	p.Pro262Leu	Missense	0	0	0	0	0	0.10
rs749979006	133533775	g.11413G>A	p.Arg282His	Missense	0.06	0	0	0	0	0
rs61710826	133533792	g.11430G>A	p.Gly288Ser	Missense	0	0	0	0	0	1.94
rs1007026512	133533813	g.11451G>A	p.Asp295Asn	Missense	0.06	0	0	0	0	0
rs533468318	133533826	g.11464C>T	p.Ala299Val	Missense	0	0	0	0	0	0.10
rs201120875	133533838	g.11476C>A	p.Thr303Asn	Missense	0	0	0	0.10	0	0

rs759707874	133537105	g.14743G>A	p.Arg337Gln	Missense	0.06	0	0	0	0	0
rs41299434	133537192	g.14830C>G	p.Ser366Cys	Missense	0	0	0	0.10	0	0
rs59981143	133537212	g.14850A>G	p.Thr373Ala	Missense	0.13	1.27	0	0.10	0	0
rs199856651	133537249	g.14887A>G	p.Lys385Arg	Missense	0	0	0	0.10	0	0
*rs55897648	133537760	g.15398G>A	p.Val389Ile	Missense	0	0	0.14	0	0.10	0
rs59656378	133537781	g.15419G>C	p.Val396Leu	Missense	0.19	0.96	0.28	0	0.10	0
rs57702102	133537848	g.15486A>G	p.Asn418Ser	Missense	0.13	0.32	0	0	0	0
rs142767992	133537865	g.15503A>G	p.Ser424Gly	Missense	0.06	0	0	0	0	0
rs199855848	133538817	g.16455G>C	p.Met445Ile	Missense	0.06	0	0	0	0	0
rs28969387	133538852	g.16490A>T	p.His457Leu	Missense	4.72	4.78	0.14	0	0	0
rs555344732	133532778	g.10416T>A	p.Tyr245Ter	Stop gain	0.06	0	0	0	0	0
rs200228968	133537104	g.14742C>T	p.Arg337Ter	Stop gain	0	0	0	0.10	0	0

Non-coding variants defining known *CYP2E1* star alleles

rsID	Location (GRCh38)	Variant alleles (NG_008383.1)	Consequence	Allele frequencies (%)					
				SSA **(n=947)**	African American/Afro-Caribbean **(n=157)**	AMR **(n=347)**	EAS **(n=504)**	EUR **(n=503)**	SAS **(n=489)**
*rs3813867	133526101	g.3739G>C	2KB Upstream	6.1	6.1	12.1	20.1	4.1	0.9
*rs2031920	133526341	g.3979C>T	2KB Upstream	0	0.6	11.7	20.2	4.1	0.9
*rs2070672	133527044	g.4682A>G	2KB Upstream	16.8	12.7	10.1	26.4	2.7	28.7
*rs2070673	133527063	g.4701A>T	2KB Upstream	22.5	31.5	66.6	54.1	81.6	61.9
*rs6413420	133527325	g.4963G>T	2KB Upstream	1.1	1.6	2.6	0.5	6.3	6.6
*rs2070676	133537633	g.15271G>C	Intron	28.0	34.4	81.4	79.5	87.2	83.0

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants

*SNVs used to define known *CYP2E1* star alleles including **1B* (rs2070676) (Mcbride et al., 1987), **2* (rs72559710), **3* (rs55897648) (Hu et al.,1997), **4* (rs6413419) (Fairbrother et al., 1998), **5A* (rs3813867, rs2031920) (Hayashi et al., 1991; Persson et al., 1993), **5B* (rs3813867, rs2031920) (Hayashi et al., 1991), *7A (rs2070676), **7B* (rs6413420, rs2070676), and **7C* (rs2070672, rs2070676) (Fairbrother et al., 1998) – definitions may be incomplete due to in adequate gene sequencing. The g.13007T>A variant defines **5A* and **6A* (Persson et al., 1993), and it was not called in this study.

Table 3: Distribution of *CYP2E1* variants across SSA populations in this study

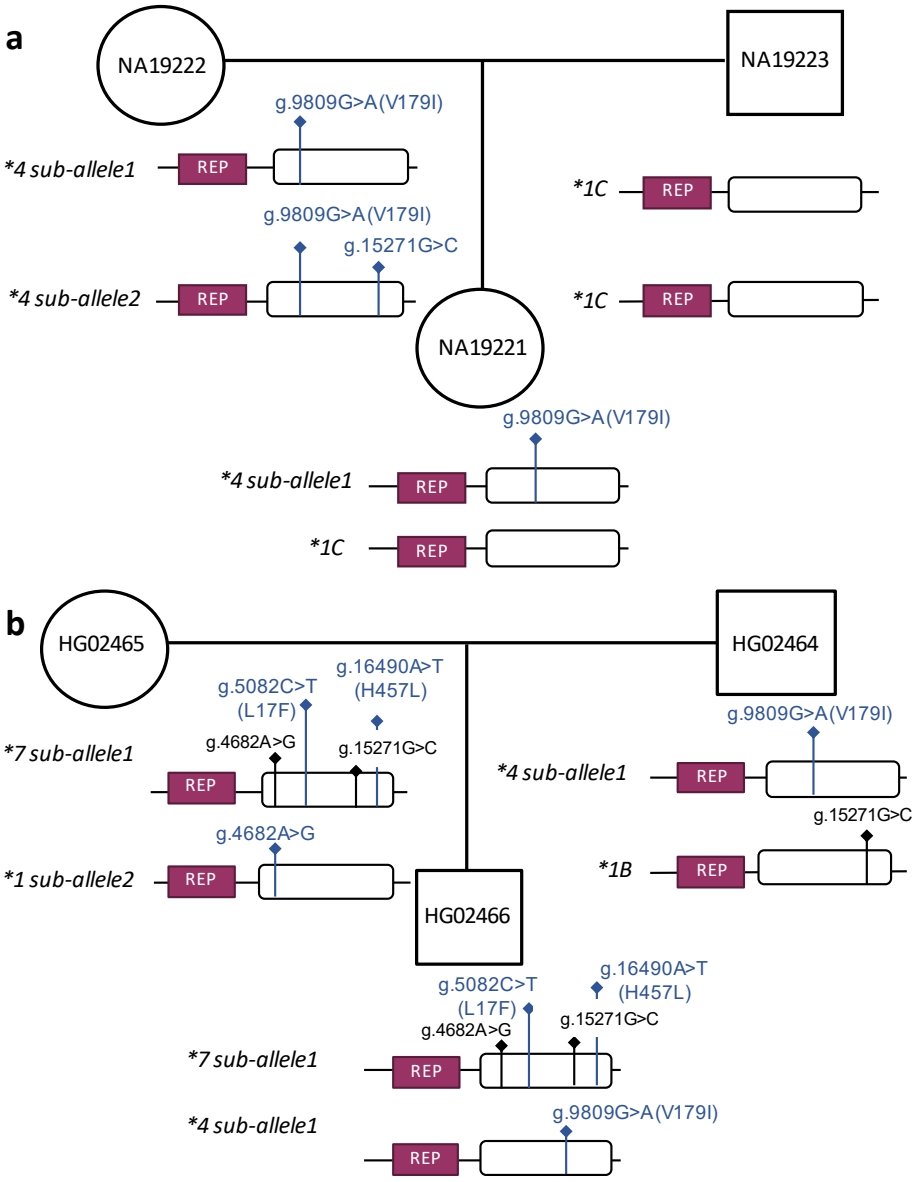
rsID	Allele frequencies (%)								
	All **(n=947)**	BFA **(n=33)**	ESN **(n=99)**	GHA **(n=26)**	GWD **(n=113)**	LWK **(n=99)**	MSL **(n=85)**	SEB **(n=200)**	YRI **(n=108)**
Non-synonymous variants									
rs367957731	0.06	0	0	0	0.4	0	0	0	0
g.5167T>C	0.06	0	0	0	0	0	0	0.3	0
*rs6413419	24.60	31.8	28.8	19.2	21.7	13.6	24.1	55.0	31.9
rs749979006	0.06	0	0	0	0	0	0	0.3	0
rs1007026512	0.06	1.5	0	0	0	0	0	0	0
rs59981143	0.13	0	0.5	0	0.4	0	0	0	0
rs59656378	0.19	0	0	0	0.4	0	0	0.3	0.5
rs57702102	0.13	3.0	0	0	0	0	0	0	0
rs142767992	0.06	0	0	0	0	0	0.6	0	0
rs199855848	0.06	0	0	0	0	0.5	0	0	0
rs28969387	4.72	9.1	3.5	5.8	6.2	2.5	5.9	4.0	6.0
rs555344732	0.06	0	0	0	0.4	0	0	0	0
Non-coding variants defining known *CYP2E1* star alleles									
*rs3813867	6.1	13.6	8.6	6.5	8.0	2.5	6.5	3.5	8.3
*rs2070672	16.8	12.1	10.6	17.7	21.7	10.6	18.2	17.8	16.2
*rs2070673	22.5	18.1	23.2	24.2	24.3	19.2	21.2	24.8	19.4
*rs6413420	1.1	0	1.5	0	3.5	1.0	1.2	0.3	0.5
*rs2070676	28.0	33.3	31.8	30.6	32.3	23.7	25.9	22.0	33.8

BFA – Burkina Faso participants; ESN – Esan in Nigeria; GWD – Gambian in Western Division, Mandinka; GHA – Ghanaian participants; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; SEB – Southeastern Bantu in South Africa; YRI – Yoruba in Ibadan

*SNVs used to define known *CYP2E1* star alleles including **1B* (rs2070676) (Mcbride et al., 1987), **2* (rs72559710), **3* (rs55897648) (Hu et al.,1997), **4* (rs6413419) (Fairbrother et al., 1998), **5A* (rs3813867, rs2031920) (Hayashi et al., 1991; Persson et al., 1993), **5B* (rs3813867, rs2031920) (Hayashi et al., 1991), *7A (rs2070676), **7B* (rs6413420, rs2070676), and **7C* (rs2070672, rs2070676) (Fairbrother et al., 1998) – definitions may be incomplete due to in adequate gene sequencing. The g.13007T>A variant defines **5A* and **6A* (Persson et al., 1993), and it was not called in this study.

CYP2E1 haplotypes characterised using trio data

Supplementary Table 3 shows a list of all the *CYP2E1* haplotypes (n=14) characterised using publicly available trio data from the 1000 Genomes Project. Of these, the most frequent (AF=29.2%) non-reference *CYP2E1* haplotype in SSA was defined by rs6413419 (g.9809G>A, p.Val179Ile) which was predicted to be deleterious (Supplementary Table 2, Supplementary Table 3). The variant is found in the current definition of *CYP2E1*4*, and our trio data analysis (involving NA19222) showed it to be on a *CYP2E1*1C* backbone (Figure 6a). Figure 6b depicts one of the potentially novel *CYP2E1* haplotypes characterised with trio data. This haplotype was identified in the GWD population, and it consisted of rs367957731 (g.5082C>T) and rs28969387 (g.16490AA>T) on a **7C* backbone, the later variant was predicted to be deleterious (Figure 6b, Supplementary Table 2, Supplementary Table 3).



NA19222 full haplotypes

Haplotype 1 (**4* sub allele 1):

2803~C>T; 5000~T>TC; 7369~G>A; 8140~T>C; 8515~G>A; 8825~C>T; 8957~T>A; 9527~G>T; 9737~G>A; **9809~G>A**; 10132~C>T; 10151~C>T; 10152~A>G; 12009~T>TGTC; 12506~AT>A; 12891~C>G; 15496~C>T; 20288~CCTTTCTTT>C; 20381~T>TTC; 20389~TTC>T

Haplotype 2 (**4* sub allele 2):

2794~C>G; 3519~T>G; 3690~C>G; 4227~T>C; 5000~T>TC; 5582~G>A; 6778~G>A; 8663~T>C; 8713~A>G; 8957~T>A; 9527~G>T; 9570~C>T; 9737~G>A; **9809~G>A**; 10132~C>T; 10151~C>T; 10152~A>G; 10830~C>T; 11667~A>G; 12007~C>CTCG; 12009~T>C; 12506~A>AT; 12891~C>G; 13360~C>A; 13501~G>A; 13671~G>A; 13935~T>C; 14048~C>T; 14213~A>G; 15005~G>T; **15271~G>C**; 15308~T>A; 15740~GTGT>G; 16147~C>T; 16210~A>G; 16234~T>C; 16287~A>G; 16302~T>C; 16648~A>G; 16822~TTTATTTTTG>T; 16989~T>C; 17213~A>C; 17633~T>G; 17752~C>T; 18259~T>G; 18385~T>G; 18665~C>A; 18749~TAAGAA>T; 18848~G>C; 19008~A>C; 19141~C>T; 19415~C>CTCTTAAT; 19416~A>G; 19528~G>A; 19534~C>T; 19545~C>A; 19760~TA>T; 20226~G>A; 20381~T>TTC; 20435~T>C; 20717~T>C; 20928~G>T; 21491~C>T; 21618~C>T; 21757~C>T

HG02465 full haplotypes

Haplotype 1 (**7* sub-allele 1):

1363~C>T; 1364~A>G; 1369~T>C; 2217~T>C; 2739~C>T; 3378~G>C; 3519~T>G; 4104~A>G; **4682~A>G**; 5000~T>TC; **5082~C>T**; 6778~G>A; 8663~T>C; 8845~T>C; 10830~C>T; 11061~C>A; 11667~A>G; 12007~C>CTCG; 12009~T>C; 12391~C>T; 12678~T>A; 12713~A>G; 12807~C>A; 12891~C>G; 12971~T>TA; 13091~C>A; 13360~C>A; 13501~G>A; 13671~G>A; 13935~T>C; 14048~C>T; 14213~A>G; 14260~T>TGGG; 15005~G>T; 15165~C>G; **15271~G>C**; 15308~T>A; 15740~GTGT>G; 15756~AT>A; 16220~G>T; 16402~C>G; **16490~A>T**; 16648~A>G; 16734~A>C; 16818~T>G; 16825~A>G; 16831~G>T; 16832~T>G; 17194~C>T; 17344~G>A; 18109~G>C; 18202~G>A; 18214~T>C; 18385~T>G; 18399~G>T; 18848~G>C; 19008~A>C; 19290~C>T; 19302~C>G; 19337~C>T; 19362~A>C; 19374~TA>T; 19383~T>C; 19386~G>A; 19760~TAA>T; 19783~A>G; 19995~C>T; 20818~T>C; 21067~G>A; 21618~C>T; 21757~C>T

Haplotype 2 (**1* sub-allele 2):

1363~C>T; 1364~A>G; 1369~T>C; 2217~T>C; 2739~C>T; 3378~G>C; 3519~T>G; 4104~A>G; **4682~A>G**; 5000~T>TC; 6394~C>T; 6778~G>A; 8663~T>C; 8845~T>C; 10830~C>T; 11061~C>A; 11667~A>G; 12007~C>CTCG; 12009~T>C; 12678~T>A; 12713~A>G; 12807~C>A; 12891~C>G; 12971~T>TA; 13091~C>A; 13360~C>A; 13501~G>A; 13671~G>A; 13935~T>C; 14048~C>T; 14162~G>A; 14213~A>G; 14260~T>TGGG; 14363~ATGGTTGGATGGATGGGTGGGTGGGTGGATGGATGGG>A; 14499~C>T; 15005~G>T; 15740~GTGT>G; 15756~AT>A; 18259~T>G; 20226~G>A; 20381~T>TTC; 20389~TTC>T; 20390~TC>*; 21491~C>T

Figure 6: Characterisation of *CYP2E1* (NG_008383.1) haplotypes from trio data. (a) Trio data showing the phasing for novel sub-alleles of *CYP2E1*4* which is defined by rs6413419 (g.9809G>A; V179I). Variant information from the child (NA19221) and the father (NA19223) enabled inference of the SNVs in phase with rs6413419 in the mother's *CYP2E1* haplotypes. (b) Trio data showing the phasing for a novel haplotype on a **7* backbone consisting of two additional variants (rs367957731, g5082C>T; rs28969387, g.16490AA>T). Variants bolded in blue indicate additional core variants whereas variants bolded in black indicate variants known to define existing *CYP2E1* star alleles.

CYP2E1 structural variants among the SSA populations

We identified a *CYP2E1* copy number > 2 at a frequency of 5.79% in SSA, this was higher in comparison to other global populations such as the EUR population (2.98%) (Figure 7a). **1Cx2* allele (Figure 7b) (AF=0.88%) was the most frequent duplication followed by a novel haplotype duplication of rs2070672 on a **1C* backbone, and **1Cx3* (triplication). The deletion of *CYP2E1* was observed at a frequency of 0.25% in SSA. The null deletion of *CYP2E1* was not observed in SSA, however, it was observed in the AMR population.

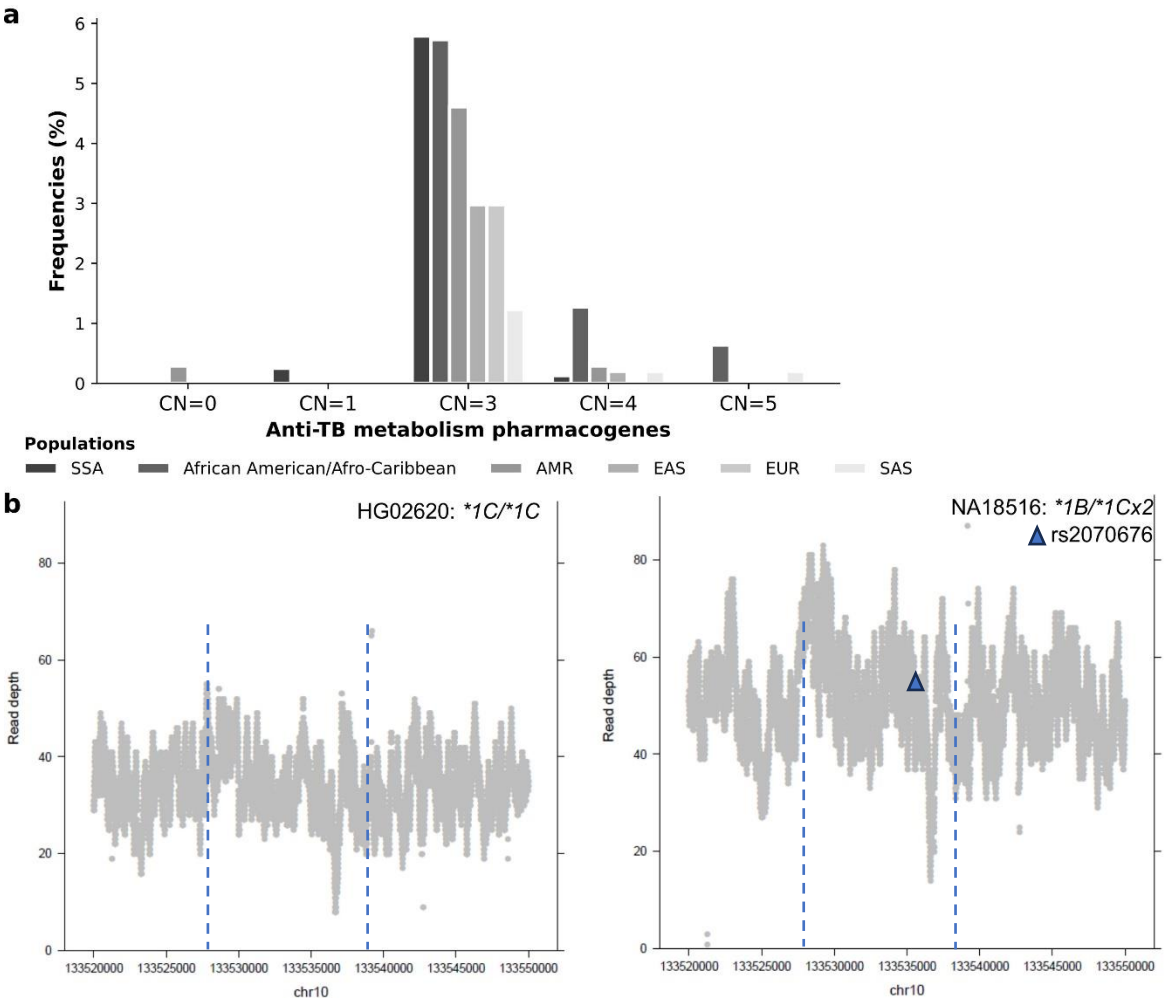


Figure 7: *CYP2E1* copy number variants identified in this study. (a) The relative frequencies of various *CYP2E1* copy numbers across Sub-Saharan African (SSA) and the global populations in this study (b) Depiction of read coverage comparing a sample with *CYP2E1* copy number=2 and one with copy number=3. The blue dotted lines indicate the extent of the *CYP2E1* genomic region (chr10: 133527363-133539123). AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants.

NAT1 and NAT2

Star allele frequencies

Six different existing *NAT1* star alleles, of varying frequencies, were called in the SSA populations (Table 4). *NAT1*4* (reference star allele — associated with slow acetylation) frequencies ranged from 32.7% (GWD) to 50.5% (LWK) (Table 5). The **4* global distribution was equally as prevalent, with SSA having the lowest frequency of 42.8% and the EUR populations having the highest average frequency of 71.3%. A significant difference was found between these two biogeographical groups ($p=4.1x10^{-43}$). In addition, the *NAT1*4* SSA frequency differed significantly ($p<2x10^{-13}$) from the frequency in the SAS (AF=59%) and AMR (AF=61.3%) populations (Table 4). The remaining *NAT1* slow acetylator star alleles (**3*, **15*, and **27*) were either rare and/or absent in SSA populations (Table 5). Similarly, many of the *NAT1* slow acetylator star alleles (**14A*, **14B*, **17*, *20, and **22*) not called in SSA showed a low prevalence in the other global populations (Table 4). In addition, *NAT1* copy number variations were not observed among the SSA populations, whereas the duplication of **10* and **4* was observed in the EAS populations, although at a low frequency (Table 4).

*NAT1*10* was the only *NAT1* rapid acetylation star allele called in the SSA populations. The *NAT1*10* frequency was lowest among the LWK (45.4%) and highest among the GHA (61.5%) and BFA (65.2%) participants (Table 5). In comparison to the other global populations included in this study, the **10* frequency in SSA (AF= 53.2%) was significantly higher ($p<0.005$) than in the African American/Afro-Caribbean (AF=42.8%), AMR (AF=35.2%), SAS (AF=33%), and EUR (AF=22.9%) populations (Table 4). No significant difference in the frequencies of the *NAT1* star alleles nor any deviations from HWE were observed between the different SSA populations.

Thirty-six unique *NAT2* star alleles were called in the SSA populations analysed in this study (Table 4, Supplementary Table 4). *NAT2*4* (reference haplotype), which is associated with rapid acetylation, was significantly less frequent in SSA (AF=9.9%), compared to the AMR (AF=30%; $p=7.9x10^{-29}$), EAS (AF=45.6%; $p=6.5x10^{-96}$), EUR (AF=23.8%; $p=3.6x10^{-21}$), and SAS (AF=19.9%; $p=1.8x10^{-12}$) populations. Among the other *NAT2* rapid acetylation star alleles, **12A* (10%) and **13A* (11.4%) were the most frequent in SSA. The **12A* frequency in SSA ranged from 3.8% in GHA to 18% in SEB whereas **13A* frequency was lowest in GHA (AF=5.8%) and highest in the MSL (AF=18.2%) populations (Table 5). These two star alleles were relatively African-specific as they were observed to be rare in other global populations, except in the African American/Afro-Caribbean populations (Table 4).

Overall, *NAT2* slow acetylation star alleles, especially **5B* (AF=16.2-25.8%) and **6A* (AF=13.3-25%), were more prevalent than the *NAT2* rapid acetylation star alleles in SSA (Table 5). However, the *NAT2*5B* frequency in SSA (AF=21.1%) was significantly lower than that in the EUR (AF=41.7%), AMR (AF=34%), and SAS (AF= 31.5%) populations. In contrast, *NAT2*5B* was significantly more frequent in SSA than in EAS (AF=3.7%; $p=7.8x10^{-34}$).

Many of the *NAT2* star alleles (~50%) called in SSA had an unknown functional activity (Supplementary Table 4). Of these star alleles, only **6O* was called in all the SSA populations analysed in the study (AF=1-7.8%), the remaining unknown function star alleles were absent in at least one of the SSA populations (Supplementary Table 5).

No significant frequency differences were observed for the *NAT2* star alleles between the various SSA populations. However, deviations from HWE were noted for some *NAT2* star alleles including **4* ($p=6.7x10^{-6}$), **12A* ($p=5.6x10^{-3}$), **13A* ($p=1.6x10^{-3}$) and **5B* ($p=3.7x10^{-12}$).

Table 4: The distribution of *NAT1* and *NAT2* star alleles in SSA compared with various global populations

Star alleles	Defining variants	Consequence	Acetylation effect[a,b]	Allele frequencies (%)*					
				SSA **(n=775)**	African American/Afro-Caribbean **(n=153)**	AMR **(n=345)**	EAS **(n=477)**	EUR **(n=482)**	SAS **(n=466)**
NAT1									
**10*	rs1057126 rs15561	3'-UTR 3'-UTR	Rapid	53.4	42.8	35.2	49.9	22.9	33.0
**11A*	rs4986988 rs4986989 rs4987076 rs4986990 rs4986783 1077_1085del(AAT)$_3$ rs15561	Intron Intron Missense (V149I) Synonymous (T153T) Missense (S214A) 3'-UTR 3'-UTR	Rapid	0	0.3	1.0	0.3	1.8	3.0
**3*	rs15561	3'-UTR	Slow	2.4	2.3	1.4	1.9	2.1	4.4
**4*	Reference	-	Slow	42.8	50.3	61.3	45.8	71.3	59.0
**14A*	rs4986782 rs1057126 rs15561	Missense (R187Q) 3'-UTR 3'-UTR	Slow	0	0.3	0.3	0	0.6	0
**14B*	rs4986782	Missense (R187Q)	Slow	0	0	0	0	0	0.3
**15*	rs5030839	Stop gain (R187*)	No enzyme activity	0.1	0.3	0.4	0	0.9	0
**17*	rs56379106	Missense (R64W)	Slow	0	0.3	0	0	0.5	0
**20*	rs146727732	Synonymous (P134P)	Slow	0	0.3	0	0	0	0
**22*	rs56172717	Missense (D251V)	Slow	0	0	0.1	0	0.5	0
**27*	rs4986992 rs4986991	Synonymous (L7L) Synonymous (S259S)	Slow	1.1	2.9	0.1	0	0	0

**16*	1091insAAA rs15561	Indel 3'-UTR	Unknown	0	0	0	1.9	0	0
**26B*	1065_1090insTAA	Indel	Unknown	0.2	0	0	0	0	0
**0*		Null deletion	Unknown	0	0	0	0	0.2	0
**10x2*		Duplication	Unknown	0	0	0	0.1	0	0.2
**3x2*		Duplication	Unknown	0	0	0	0	0	0.1
**4x2*		Duplication	Unknown	0	0	0	0.1	0	0

NAT2

Star alleles	Defining variants	Consequence(s)	Acetylation effect[b]	Allele frequencies (%)*					
				SSA **(n=792)**	African American/Afro-Caribbean **(n=157)**	AMR **(n=346)**	EAS **(n=504)**	EUR **(n=503)**	SAS **(n=488)**
**4*	Reference	-	Rapid	9.9	13.1	30.0	45.6	23.8	19.9
**12A*	rs1208	Missense (K268R)	Rapid	10.0	4.5	1.2	0.2	0.3	1.7
**12B*	rs1041983 rs1208	Synonymous (Y94Y) Missense (K268R)	Rapid	1.2	1.0	0.3	0	0.1	0
**12C*	rs1799929 rs1208	Synonymous (L161L) Missense (K268R)	Rapid	0.1	0	0	0	0	0.1
**13A*	rs1041983	Synonymous (Y94Y)	Rapid	11.4	10.8	0.7	0.5	0	0
**5A*	rs1801280 rs1799929	Missense (I114T) Synonymous (L161L)	Slow	0.4	0.3	0.6	0	1.7	0.3
**5B*	rs1801280 rs1799929 rs1208	Missense (I114T) Synonymous (L161L) Missense (K268R)	Slow	21.2	24.2	34.0	3.7	41.7	31.5
**5C*	rs1801280 rs1208	Missense (I114T) Missense (K268R)	Slow	6.1	4.1	1.2	0.1	1.6	2.6
**5D*	rs1801280	Missense (I114T)	Slow	0.3	0	0.9	0	0	0
**5E*	rs1801280 rs1799930	Missense (I114T) Missense (R197Q)	Slow	0.1	0	0	0	0	0
**6A*	rs1041983 rs1799930	Synonymous (Y94Y) Missense (R197Q)	Slow	16.9	20.7	16.8	25.9	27.8	35.3

**6B*	rs1799930	Missense (R197Q)	Slow	0	0	0.1	0	0.2	0
**6C*	rs1041983 rs1799930 rs1208	Synonymous (Y94Y) Missense (R197Q) Missense (K268R)	Slow	0	0	0	0	0.1	0.3
**7B*	rs1041983 rs1799931	Synonymous (Y94Y) Missense (G286E)	Slow - substrate dependent	1.9	4.1	13.6	23.8	2.4	7.8
**10*	rs72554617	Missense (E167K)	Slow	0	0	0	0.1	0	0
**14A*	rs1801279	Missense (R64Q)	Slow	3.0	1.6	0.1	0	0	0
**14B*	rs1801279 rs1041983	Missense (R64Q) Synonymous (Y94Y)	Slow	6.3	4.5	0.1	0	0.1	0
**14D*	rs1801279 rs1041983 rs1799930	Missense (R64Q) Synonymous (Y94Y) Missense (R197Q)	Slow	0.4	0	0	0	0	0
**14E*	rs1801279 rs1208	Missense (R64Q) Missense (K268R)	Slow	0.1	0	0	0	0	0
**14G*	rs1801279 rs1041983 rs1208	Missense (R64Q) Synonymous (Y94Y) Missense (K268R)	Slow	0	0.3	0	0	0	0

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants.

UTR – untranslated region

[a]Acetylation effect based on those stated in the literature (Loktionov et al. (2002), and Wang et al. (2011)) and [b]Database of NATs (http://nat.mbg.duth.gr/) and PharmGKB (https://www.pharmgkb.org/vip/PA166170337).

*Allele frequencies excluded novel haplotypes and individuals with unresolved diplotypes.

Table 5: The distribution of *NAT1* and *NAT2* star alleles variants across SSA populations in this study

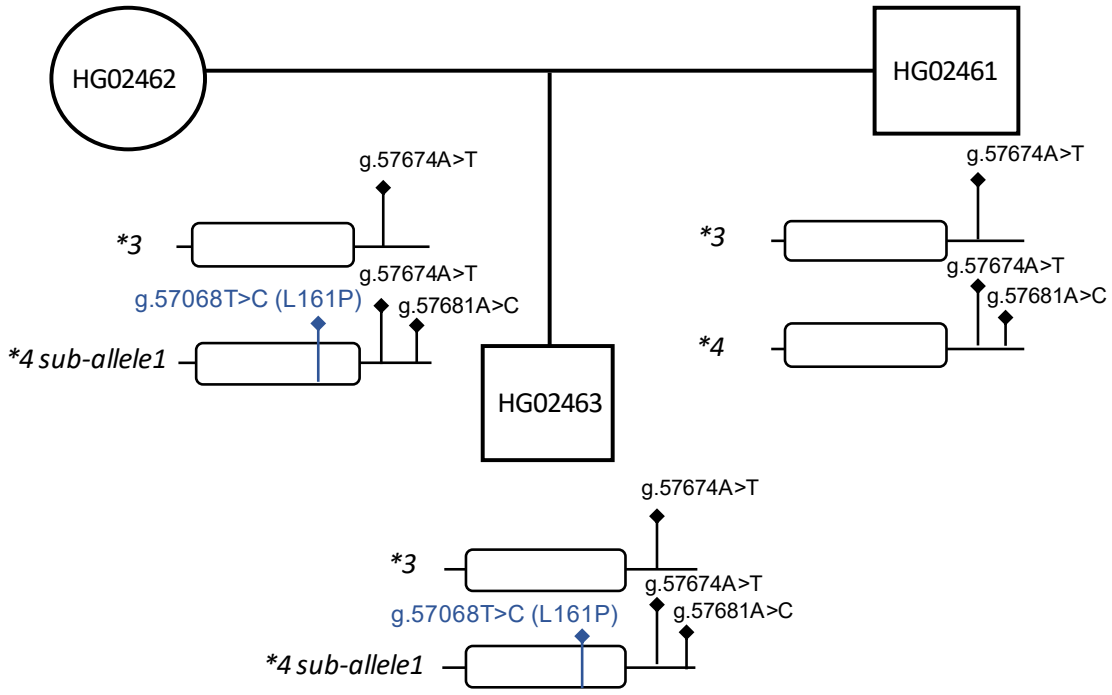
Star alleles	Acetylation effect[a,b]	Allele frequencies (%)*								
		All (n=775)	BFA (n=33)	ESN (n=96)	GWD (n=112)	GHA (n=26)	LWK (n=97)	MSL (n=81)	SEB (n=195)	YRI (n=106)
					NAT1					
**10*	Rapid	53.4	65.2	51.0	58.3	61.5	45.4	59.3	54.1	47.2
**3*	Slow	2.4	7.6	1.6	1.8	3.8	2.6	3.7	1.8	1.9
**4*	Slow	42.8	27.3	46.4	37.7	32.7	50.5	36.4	42.6	50.0
**15*	Slow	0.1	0	0	0.5	0	0	0	0.3	0
**27*	Slow	1.1	0	1.0	0.5	1.9	1.5	0.6	1.3	0.9
**26B*	Unknown	0.2	0	0	1.3	0	0	0	0	0
Star alleles	Acetylation effect[a,b]	Allele frequencies (%)*								
		All (n=792)	BFA (n=32)	ESN (n=98)	GWD (n=112)	GHA (n=26)	LWK (n=99)	MSL (n=85)	SEB (n=200)	YRI (n=108)
					NAT2					
**4*	Rapid	9.9	4.7	7.1	11.6	9.6	7.6	14.1	9.0	12.5
**12A*	Rapid	10.0	10.9	7.1	4.9	3.8	6.6	7.1	18.0	6.5
**12B*	Rapid	1.2	0	0.5	0	0	0	1.2	3.8	0.5
**12C*	Rapid	0.1	0	0	0.4	0	0	0	0	0
**13A*	Rapid	11.4	15.6	14.2	8.5	5.8	9.6	18.2	8.0	16.7
**5A*	Slow	0.4	0	0	1.3	0	0	0	0.8	0
**5B*	Slow	21.2	21.9	18.8	25.9	23.0	17.0	31.3	16.2	18.8
**5C*	Slow	6.1	6.3	8.1	4.9	5.8	8.0	3.0	6.9	4.7
**5D*	Slow	0.3	1.6	0	0.4	0	0.5	0	0.5	0
**5E*	Slow	0.1	0	0.5	0	0	0	0	0	0
**6A*	Slow	16.9	14.1	21.3	13.4	25.0	14.3	22.7	14.8	16.7
**7B*	Slow (substrate dependent)	1.9	1.6	1.0	0.9	5.8	7.5	1.0	4.2	4.1
**14A*	Slow	3.0	0	1.5	5.8	7.7	3.3	1.0	3.2	2.9
**14B*	Slow	6.3	12.5	10.2	7.6	7.7	3.8	6.1	6.9	3.5
**14D*	Slow	0.4	0	1.0	0.9	0	0	0	0.9	0
**14E*	Slow	0.1	0	0	0	0	0	0.5	0	0

BFA – Burkina Faso participants; ESN – Esan in Nigeria; GWD – Gambian in Western Division, Mandinka; GHA – Ghanaian participants; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; SEB – Southeastern Bantu in South Africa ; YRI – Yoruba in Ibadan.
[a]Acetylation effect based on literature (Loktionov et al. (2002) and Wang et al. (2011)), and [b]Database of NATs (http://nat.mbg.duth.gr/) and PharmGKB (https://www.pharmgkb.org/vip/PA166170337).
*Allele frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes.

Potentially novel *NAT1* and *NAT2* haplotypes and diplotypes

There were 16 unique potentially novel *NAT1* haplotypes found in the SSA study populations (Table 6). In addition, StellarPGx flagged five other potential novel *NAT1* star alleles, but these occurred in unresolved diplotypes (see Discussion). Trio data corresponding to some of the 1000G samples enabled further characterisation of novel haplotypes 1-9 (i.e. ascertaining all

variants in phase with the core variants) (Supplementary Table 6). In particular novel haplotype 1, defined by rs554793519 (g.57068T>C) on a **4* backbone, occurred in an initially unresolved diplotype according to StellarPGx, but this was resolved following consideration of trio information (Figure 8). Among the variants defining the novel star alleles rs554793519, rs529845573, rs138061602, and g.57013~C>T (novel variant) were predicted to be deleterious (Supplementary Table 2). Additionally, rs554793519, rs529845573, rs138061602 (missense variants) were predicted to destabilise NAT1 ($\Delta\Delta G^{Stability}$ = -0.58 – -1.72kcal/mol).



HG02462 full haplotypes

Haplotype 1 (**4 sub-allele*):	Haplotype 2 (**3*):
57068~T>C; 57674~A>T; 57681~A>C; 57777~T>G	**57674~A>T**; 58302~C>T

Figure 8: Characterisation of a potentially novel *NAT1* haplotype using inheritance information. As initially predicted using StellarPGx, the novel haplotype has a *NAT1*4* background. The novel haplotype has **4* background allele (consisting of two 3'UTR variants) and an additional missense variant (rs554793519~g.57068T>C). The male child inherited the novel haplotype from their mother, and inherited **3* from their father. Variants bolded in blue indicate additional core variants whereas variants bolded in black indicate variants known to define existing *NAT1* star alleles. UTR – untranslated region

Regarding *NAT2*, one potentially novel haplotype was identified in the ESN population that had rs561124342 on a **6A* backbone (Table 6, Supplementary Table 6). The additional missense variant (rs561124342) was predicted to be deleterious (Supplementary Table 2). Additionally, the core missense variants defining the potentially novel haplotype was predicted to impact the stability of NAT2 ($\Delta\Delta G^{Stability}$ = -1.7 kcal/mol).

Table 6: Possible novel *NAT1* and *NAT2* haplotypes and unresolved diplotypes in the SSA populations

#	Background allele	Additional core variant(s)	Consequences	Allele count	Population
Resolved haplotypes (confirmed with trio data)					
NAT1					
1	**4*	rs554793519~g.57068T>C	Missense (L161P)	1	GWD[a]
2	**4*	rs201916736~g.57277G>T	Missense (V231F)	1	YRI[a]
3	**4*	g.57013~C>T	Stop gain (Q143*)	1	SEB[c]
4	**4*	rs4986992~g.56607T>G	Synonymous (L7L)	3	ESN[a], Nigeria[e]◆
5	**4*	g.57669delAAT	3'-UTR	3	SEB[b,d]
6	**4*	g.57669insAAT	3'-UTR	1	ESN[a]
7	**4*	g.57669insAAT	3'-UTR	1	ESN[a]
8	**3*	g.57669ins(AAT)$_3$	3'-UTR	1	GWD[a]
9	**10*	rs111848753~g.56684G>A	Missense (R33Q)	3	SEB[b,c,d]
10	**10*	rs138061602~g.56777G>A	Missense (R64Q)	1	ESN[a]
11	**10*	rs529845573~g.57025C>T	Missense (P147S)	1	GWD[a]
12	**10*	rs1389255907~g.57437G>A	Missense (G284D)	1	SEB[b]
13	**10*	rs4986991~g.57363T>C	Synonymous (S259S)	1	LWK[a]
14	**10*	rs15561~g.57681A>C	3'-UTR	3	SEB[c], South Africa[e]
15	**10*	g.57666delAATAATAAA	3'-UTR	3	LWK[a]
16	**10*	g.57669insAAT	3'-UTR	1	YRI[a]
NAT2					
1	**6A*	rs561124342~g.14007G>C	Missense (G83A)	1	ESN[a]
Unresolved novel diplotypes					
#	Background allele	Additional core variant(s)	Consequences	Count	Population
NAT1					
1	**4/*10*	rs370302501~g.57038G>A	Missense (R151H)	1	MSL[a]
2	**4/*10*	rs370859724~g.56901G>A	Missense (M105I)	1	South Africa[e]
3	**4/*10*	g.57669ins(AAT)$_3$	3'-UTR	1	MSL[a]

4	**4/*10*	g.57669insAAT	3'-UTR	1	MSL[a]
5	**10/*10*	rs548004925~g.56972A>G+ rs932721238~g.57671delT(A)$_5$+ rs15561~g.57681A>C	Missense (Y129C) 3'-UTR, 3'-UTR	1	MSL[a]
6	**10/*10*	g.57669insAAT + rs15561~g.57681A>C	3'-UTR, 3'-UTR	1	YRI[a]
		NAT2			
1	**5C/*11A*	rs79050330~14337~C>T	Missense (T193M)	1	GWD[a]
2	**7B/*12A*	rs146405047~14113~T>G	Missense (N118K)	1	BFA[b]

[a]1000G, [b]AWI-Gen, [c]CBRL, [d]SAHGP, [e]SGDP

◆Haplotypes of two samples from the ESN population (1000G) and a participant from Nigeria (SGDP) could not be confirmed using trio data

BFA – Burkina Faso participants, ESN – Esan in Nigeria; GWD – Gambian in Western Division, Mandinka; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; SEB – Southeastern Bantu in South Africa; YRI – Yoruba in Ibadan

UTR – untranslated region

NAT1 and NAT2 phenotype predictions

The NAT1 slow acetylation phenotype (mainly accounted for by **4/*4*) was significantly less frequent in SSA (20.8%) especially in comparison to AMR (39%), SAS (38.8%) and EUR (58.8%) populations (Figure 9 and Supplementary Table 7). The NAT1 intermediate acetylator phenotype was the most predominant phenotype in SSA (51.1%) and other global populations (Figure 9). The relatively high prevalence among the SSA populations was significantly higher than in the AMR (49.6%, $p=3.9x10^{-2}$), and EUR (35.7%, $p=2.0x10^{-6}$) populations. The observed NAT1 intermediate acetylator phenotype was mainly attributed to the **4/*10* diplotype, which had a high frequency among the African American/Afro-Caribbean participants (49.3%), followed by SSA (47.1%) participants (Supplementary Table 7). The frequency of the NAT1 rapid acetylator phenotype in SSA was 27.9% and was significantly higher in most of the global populations apart from the EAS population (p=1.0). The only diplotype associated with this phenotype in the various global populations in this study was **10/*10* (Supplementary Table 7).

Between 38-62.7% of individuals across the global populations were predicted to be NAT2 slow acetylators (Figure 9). In the SSA populations, 42% of the participants were predicted to be NAT2 slow acetylators, significantly lower compared to the 58.3% in EUR ($p=8.7x10^{-7}$) and the 62.7% in SAS ($p=1.3x10^{-10}$) populations. The **5B/*6A* diplotype contributed most to the NAT2 slow acetylator phenotype in each of the populations, except in the AMR were the **5A/*14B* diplotype was the main contributor (Supplementary Table 8). The NAT2 intermediate acetylator phenotype was observed highest among the SSA populations (45%). The frequency of this phenotype was significantly higher in SSA when compared to the other global populations, except for the African American/Afro-Caribbean and AMR populations. NAT2 diplotypes which accounted for this phenotype were observed at relatively low frequencies individually in SSA (Supplementary Table 8). The NAT2 rapid acetylator phenotype was only observed in 13% of the SSA participants. In most of the global populations the NAT2 rapid acetylator was less frequent, with African American/Afro-Caribbean (10%), EUR (6.8%), SAS (6.0%) populations having lower frequencies than what was observed in SSA for this phenotype. The only exception was in the EAS populations, where 31% of individuals were NAT2 rapid acetylators. This was significantly higher when compared to SSA ($p=9.7x10^{-12}$). In SSA the phenotype was mainly attributable to the **4/*13A* diplotype, whereas the **4/*4* diplotype accounted most for this phenotype in AMR, EAS, EUR, and SAS populations (Supplementary Table 8).

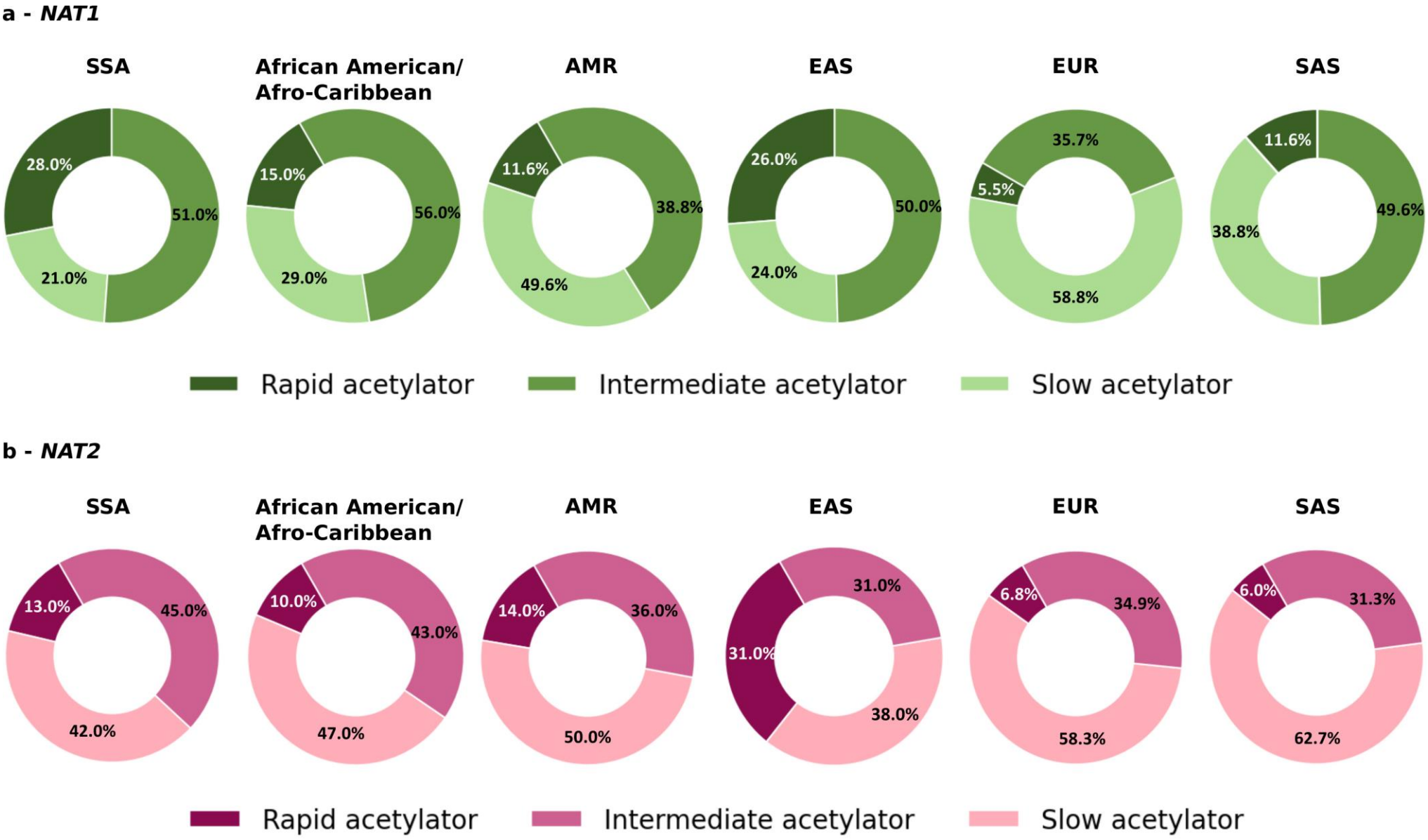


Figure 9: Distribution of NAT1 and NAT2 phenotype predictions across Sub-Saharan Africa (SSA) compared to other global populations. NAT1 (a) and NAT2 (b) intermediate and slow acetylator were relatively common across the global populations. NAT1 and NAT2 rapid acetylator varied. The SSA populations had a higher frequency of NAT1 rapid acetylators (28%), while lower frequencies were reported for NAT2 rapid acetylators (13%). AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants.

GSTM1 and GSTT1

Star allele frequencies

The *GSTM1*0* allele (gene deletion) (Figure 10) was common across the various SSA populations ranging from 43.7% in SEB to 57.4% in the ESN (Table 8) population. A higher prevalence of *GSTM1*0* was observed among the global populations, especially in EUR (73.5%) and EAS (75.3%) populations (Table 7). Both these frequencies were significantly higher compared to the average frequency in SSA (AF=49.2%).

*GSTM1*A* (associated with normal GSTM1 enzyme activity) was relatively frequent in the SSA populations (AF=38.5-52.1%) (Table 7). When comparing to the other global populations, the **A* distribution was similar between SSA and the African American/Afro-Caribbean population (Table 7). When compared to the remaining global populations, SSA had a significantly higher **A* frequency (Table 7). Both, **Ax2* and **Bx2* were relatively rare, with **Bx2* only being called in the SAS population.

Both *GSTT1*0* and *GSTT1*A* occurred at high frequencies of 65.6% and 34.4%, respectively, in SSA (Table 7). *GSTT1*0* was most prevalent especially in BFA (AF=70%), GWD (AF=72.3%), and ESN (AF=73.5%) (Table 8). There were no significant differences in the **0* and **A* frequency among the SSA populations nor were there any observed deviations from HWE. The high *GSTT1*0* frequency was observed in both the SSA and the other comparative global populations (Table 7). However, the **0* frequency in SSA differed significantly from the AMR (AF=49.1%), EUR (AF=40.2%), and SAS (AF=42.1%) populations.

The *GSTT1*A* frequency in SSA (34.4%) was significantly lower when compared to the frequency among the AMR (AF=50.8%, $p=1.73x10^{-11}$), EUR (AF=59.6%, $p=4.79x10^{-32}$) and SAS (AF=57.8%, $p=2.21x10^{-27}$) populations. The **Ax2* duplication was only called in the AMR, EUR, SAS populations at low frequencies (AF=0.5%) and was absent in SSA (Table 7).

The null deletion of both *GSTM1* and *GSTT1* was observed in 7.3% of the participants in SSA (Table 7). The deletion was observed at a slightly higher frequency compared to the presence of both genes (5.4%). (Table 7).

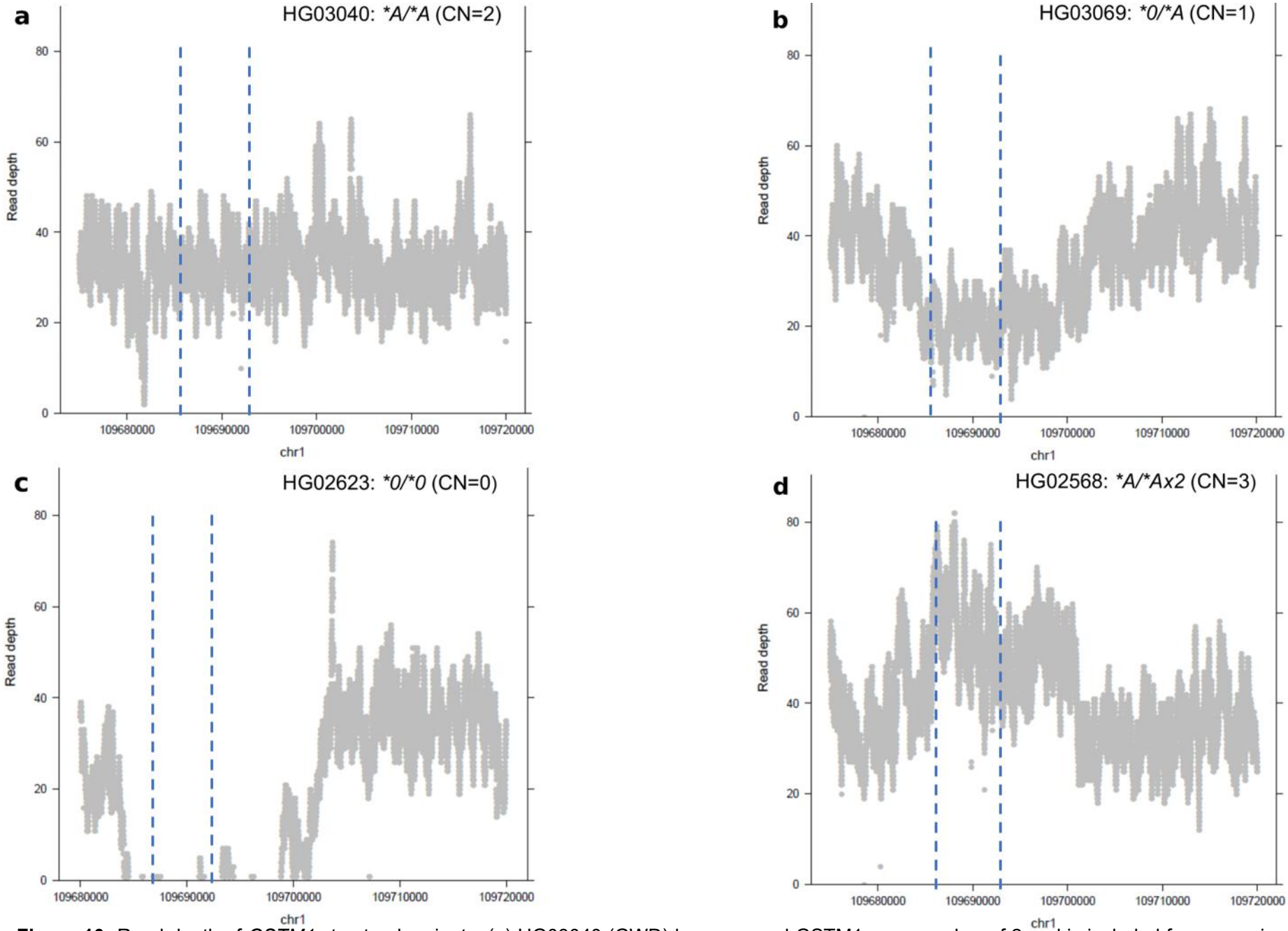


Figure 10: Read depth of *GSTM1* structural variants. (a) HG03040 (GWD) has a normal GSTM1 copy number of 2 and is included for comparison. (b) HG03069 (MSL), has a heterozygous deletion. (c) HG02623 represents an GWD individual with homozygous *GSTM1* deletion. (d) HG02568 (GWD) has a *GSTM1* duplication subsequently the individual has three *GSTM1* copies. The dotted blue lines indicate the *GSTM1* region (chr1:109687817-109693745).

Table 7: The distribution of *GSTM1* and *GSTT1* star alleles in SSA compared with various global populations

Star alleles	Defining variants	Consequence	Detoxification activity[a]	Allele frequency (%)*					
				SSA **(n=779)**	African American/ Afro-Caribbean **(n=154)**	AMR **(n=343)**	EAS **(n=504)**	EUR **(n=495)**	SAS **(n=483)**
GSTM1									
**0*	-	Gene deletion	None	49.2	50.5	69.8	75.3	73.5	57.7
**A*	Reference	-	Normal	46.4	45.0	18.7	9.1	17.7	32.9
**B*	rs1065411	Missense (K173N)	Normal	4.1	4.6	11.4	15.6	8.6	9.0
**Ax2*		Gene duplication	Increased	0.3	0	0.1	0	0.2	0.3
**Bx2*		Gene duplication	Unknown	0	0	0	0	0	0.1

Star alleles	Defining variants	Consequence	Detoxification activity[a]	Allele frequency (%)*					
				SSA **(n=649)**	African American/ Afro-Caribbean **(n=139)**	AMR **(n=343)**	EAS **(n=498)**	EUR **(n=500)**	SAS **(n=483)**
GSTT1									
**0*	-	Gene deletion	None	65.6	58.1	49.1	68.7	40.2	42.1
**A*	Reference	-	Normal	34.4	41.9	50.8	31.3	59.6	57.8
**Ax2*		Gene duplication	Unknown	0	0	0.1	0	0.2	0.1

	Diplotype frequencies (%)					
	SSA **(n=794)**	African American/Afro-Caribbean **(n=157)**	AMR **(n=347)**	EAS **(n=504)**	EUR **(n=503)**	SAS **(n=489)**
GSTM1-GSTT1						
GSTM1-/- | GSTT1-/-	7.3	5.1	11.2	26.2	8.9	5.3
GSTM1+/+ | GSTT1+/+	5.4	8.9	2.3	0	3.0	6.1

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants

*Allele frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

[a]Enzyme activity was defined based on those reviewed in Townsend & Tew (2003)

Table 8: The distribution of *GSTM1* and *GSTT1* star alleles variants across SSA populations in this study

Star alleles	Detoxification activity [a,b]	Allele frequency (%)*								
		All **(n=779)**	BFA **(n=33)**	ESN **(n=98)**	GWD **(n=111)**	GHA **(n=25)**	LWK **(n=96)**	MSL **(n=81)**	SEB **(n=199)**	YRI **(n=108)**
GSTM1										
**0*	None	49.2	56.9	57.4	50.0	50.0	55.2	48.4	45.3	43.7
**A*	Normal	46.4	38.5	39.5	45.9	46.0	42.7	48.4	47.9	51.2
**B*	Normal	4.1	4.6	3.1	3.6	2.0	2.1	2.5	6.5	4.7
**Ax2*	Increased	0.3	0	0	0.5	2.0	0	0.6	0.3	0.5
Star alleles	Detoxification activity [a,b]	Allele frequency (%)*								
		All **(n=649)**	BFA **(n=29)**	ESN **(n=81)**	GWD **(n=96)**	GHA **(n=46)**	LWK **(n=81)**	MSL **(n=73)**	SEB **(n=146)**	YRI **(n=92)**
GSTT1										
*0	None	65.6	70.0	73.5	72.3	60.9	63.4	57.5	62.3	67.8
*A	Normal	34.4	31.0	26.5	27.7	39.1	36.6	42.5	37.7	32.2

BFA – Burkina Faso participants; ESN – Esan in Nigeria; GWD – Gambian in Western Division, Mandinka; GHA – Ghanaian participants; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; SEB – Southeastern Bantu in South Africa; YRI – Yoruba in Ibadan

*Allele frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

[a]Detoxification activity was defined based on those reviewed in Townsend & Tew (2003)

Potentially novel *GSTM1* and *GSTT1* star alleles

Nine novel *GSTM1* haplotypes were identified in the SSA populations, with eight novel haplotypes having **A* as a background allele (Table 9). Three missense variants, rs375154572, rs184653774, and rs533860247 were predicted to be deleterious (Supplementary Table 2) and to destabilise GSTM1 ($\Delta\Delta G^{Stability}$ = -0.55kcal/mol, -0.5 kcal/mol, and -0.89 kcal/mol; respectively).

Ten novel *GSTT1* haplotypes were identified in the SSA populations (Table 9). The *GSTT1* variants could not be annotated using VEP, due to the gene being in a scaffold region and therefore none of the *in-silico* prediction tools could be used. However, DynaMut2 predicted GSTT1 to be destabilised by rs2266637 ($\Delta\Delta G^{Stability}$ = -0.83kcal/mol) and rs141759372 ($\Delta\Delta G^{Stability}$ = -1.82kcal/mol).

Table 9: Possible novel *GSTM1* and *GSTT1* haplotypes in the SSA populations

#	Background allele	Additional core variant(s)	Consequence(s)	Allele count	Population/dataset
			GSTM1		
1	**A*	rs449856~g.10471T>A	Missense (S210T)	11	MSL, ESN, LWK, YRI[a] SEB[b,c]
2	**A*	rs375154572~ g.7704C>T	Missense (R168C)	6	MSL[a]
3	**A*	rs138440339~ g.7677C>G + rs1065411~ g.7721G>C	Missense (L159F) Missense (K173N)	3	GWD, MSL[a]
4	**A*	rs146668816~g.6330~C>G	Missense (N85K)	3	GWD, LWK[a]
5	**A*	rs184653774~g.5105~C>A	Missense (D9E)	2	LWK[a], South Sudan[e]
6	**A*	rs533860247~g.10480~G>A	Missense (A213T)	1	ESN[a]
7	**A*	rs146668816~g.6330~C>G + rs1065411~g.7721~G>C	Missense (N85K) Missense (K173N)	1	MSL[a]
8	**A*	rs138440339~g.7677~C>G + rs1065411~g.7721~G>C (CN=4)	Missense (L159F) Missense (K173N)	1	GHA[b]
9	**B*	rs1065411~g.7721~G>C rs449856~g.10471~T>A	Missense (K173N) Missense (S210T)	1	BFA[b]
			GSTT1		
1	**A*	rs2266637~g.271020C>T + rs8140585~g.277188C>T	Missense (V169I) Missense (A61T)	175	ESN, GWD, LWK, MSL[a] , BFA, GHA[b,] SEB[b,c,d], Nigeria[e]
2	**A*	rs2266637~g.271020C>T	Missense (V169I)	35	ESN, GWD, LWK, MSL, YRI[a],GHA[b] , SEB[b,c], Gambian[e]
3	**A*	rs8140585~g.277188C>T	Missense (A61T)	54	ESN, GWD, LWK, MSL,YRI[a], BFA, GHA[b], SEB[b,c,d], South Africa, Kenya[e]
4	**A*	rs141759372~g.273586A>G	Missense (W101R)	8	ESN, MSL, YRI[a], SEB[b]
5	**A*	rs144686326~270759-G>A + rs141759372~273586-A>G	Stop gain (R188*) Missense (W101R)	1	ESN[a]
6	**A*	rs144686326~270759-G>A	Stop gain (R188*)	2	MSL[a], BFA[b]
7	**A*	rs199521920~275948-C>T	Missense (D43N)	1	LWK[a]

8	*A	rs1601976046~g.270731C>T+ rs8140585~277188C>T	Missense (R197H) Missense (A61T)	9	SEB[b,c,d]
9	*A	rs1601976046~g.270731C>T	Missense (R197H)	3	SEB[b,c]
10	*A	rs1601976046~g.270731C>T+ rs2266637~g.271020C>T+ rs8140585~g.277188C>T	Missense (R197H) Missense (V169I) Missense (A61T)	2	SEB[c,d]

[a]1000G, [b]AWI-Gen, [c]CBRL, [d]SAHGP, [e]SGDP

BFA – Burkina Faso participants; ESN – Esan in Nigeria; GWD – Gambian in Western Division, Mandinka; GHA – Ghanaian participants; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; SEB – Southeastern Bantu in South Africa; YRI – Yoruba in Ibadan

GSTM1 and GSTT1 predicted phenotypes

In the SSA populations, 24% of participants had a homozygous *GSTM1* deletion, resulting in no production of the enzyme. This frequency was significantly different from that observed in the EUR (53%, $p=4.7x10^{-23}$) and EAS (56%, $p=1.3x10^{-28}$) (Figure 11). The reduced GSTM1 activity phenotype was predominant among the various global populations, especially in SSA (49%) and SAS (48%). A significant difference between SSA and EAS (38%) was observed. Both the **0/*A* and **0/*B* diplotype contributed to this phenotype. In the SSA populations the **0/*A* was the most predominant diplotype, whereas the **0/*B* diplotype accounted more to the reduced GSTM1 activity in the EUR population (Supplementary Table 9).

The absence of GSTT1 activity (inferred from the **0/*0* diplotype) was largely prevalent among SSA (43%), African American/Afro-Caribbean (31%), and EAS (47%) (Figure 11, Supplementary Table 10). The frequency of this phenotype was significantly higher in SSA compared to the frequencies observed in the AMR (24%, $p=2.5x10^{-7}$), EUR (18%, $p=5.4x10^{-17}$), and SAS (16%, $p=2.2x10^{-19}$) populations (Figure 11). The reduced GSTT1 activity phenotype was the most prevalent GSTT1 phenotype in the SSA populations (45%). The high prevalence of this phenotype was observed in other global populations such as in the AMR, SAS and EUR populations (Figure 11). However, there were no significant differences in the frequency between the various global populations. The SSA, African American/Afro-Caribbean and EAS populations had a lower frequency of the normal GSTT1 phenotype, compared to the remaining three populations (Figure 11). The distribution of this phenotype was significantly lower in SSA (12%) compared to the AMR (25%), SAS (32%), and EUR (37%) populations (Figure 11).

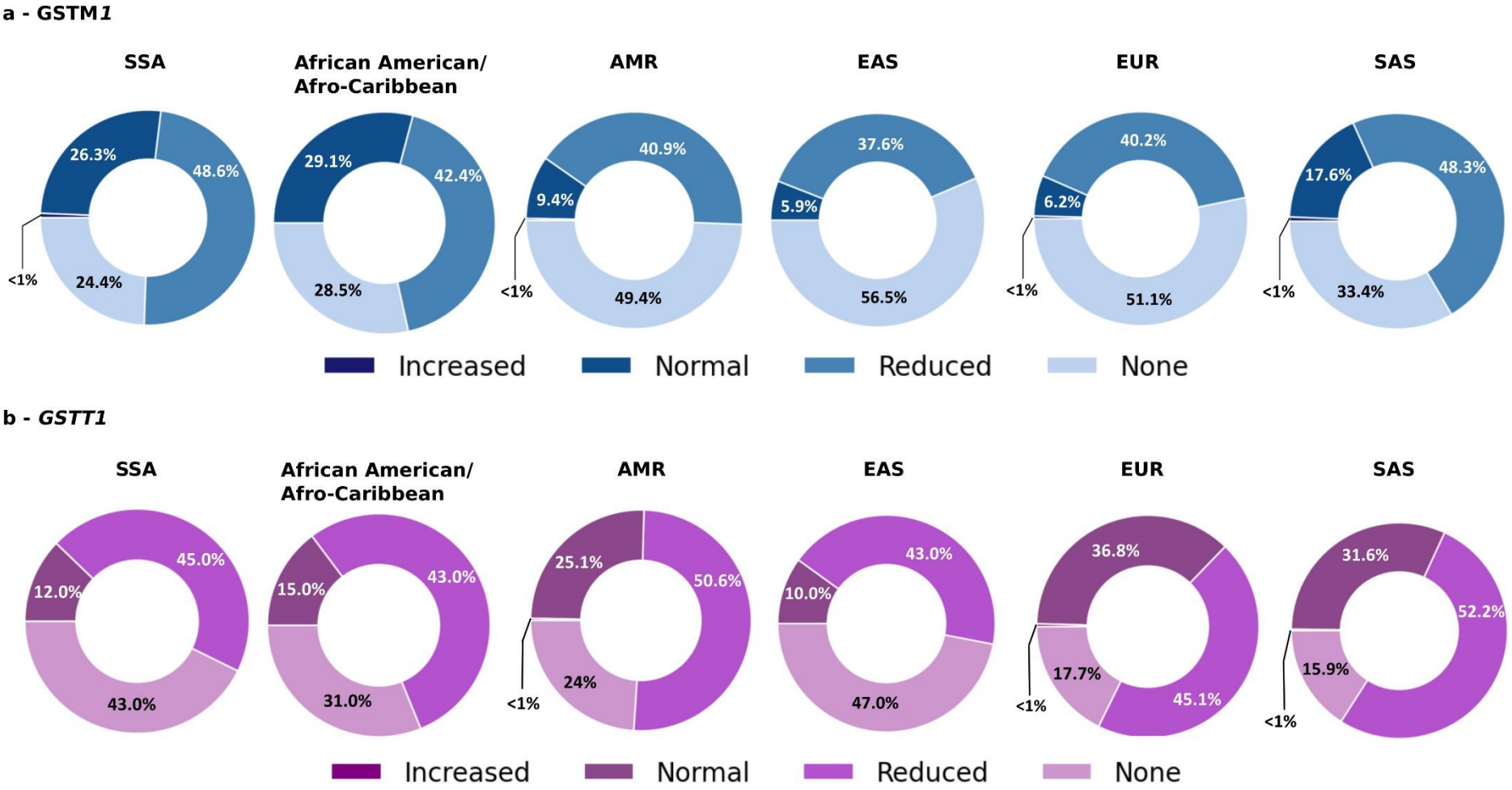


Figure 11: Distribution of GSTM1 (a) and GSTT1 (b) phenotype predictions across SSA compared to other global populations. In SSA, approximately 45% of individuals were predicted to have the reduced GSTM1 activity phenotype due to a high prevalence of heterozygous *GSTT1* deletions. The frequency of normal and the absence of GSTM1 varied in the different populations. Only the increased GSTT1 activity phenotype was relatively rare across the six population groups. SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants

Discussion

The high-TB-burden in Africa emphasises a need for effective preventative measures and treatments. However, with the high occurrence of ADRs to anti-TB drugs, in particular AT-DILI (Chen et al., 2015), it is necessary to optimise drug dosages to improve drug safety and efficacy. One major obstacle in adopting this precision medicine approach in Africa is the lack of pharmacogenomic profiling of function altering star alleles influencing drug response. While there have been some previous efforts in characterising star alleles in the context of anti-TB drug response in Africa (Ebeshi et al., 2011; Heathfield et al., 2014), many of these studies either consisted of smaller study population sizes or only analysed variation in single pharmacogenes, therefore there is a paucity of information on the extent of pharmacogenomic variation that may influence anti-TB drug response across Africa. To improve on previous Africa-based anti-TB pharmacogenomics studies, we analysed 794 high coverage genomes of participants from various SSA ethnolinguistic groups, to establish a more comprehensive insight into the pharmacogenomic diversity in five anti-TB pharmacogenes (*CYP2E1*, *NAT1, NAT2, GSTM1,* and *GSTT1*) in the African context. Furthermore, we leveraged publicly available WGS data to compare the pharmacogenomic variation landscape in these genes in SSA with that of the other global biogeographical groups.

Before conducting the main study analysis, we performed a benchmarking evaluation of StellarPGx, which has facilitated star allele analysis in previous pharmacogenomic studies (Samarasinghe et al., 2023) — especially for *CYP* genes — to ascertain its accuracy in diplotyping *CYP2E1*, *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* using WGS data. StellarPGx star allele calls were largely concordant (haplotype concordance > 80%) with GeT-RM 'gold standard' calls (Pratt et al., 2016) for *NAT2, GSTM1,* and *GSTT1,* but lower concordance was observed for *NAT1* and *CYP2E1*. Upon further investigation, the low concordance for *NAT1* could possibly be attributed to the lack of genotyping of non-coding variants (e.g., rs15561 and rs1057126) that define many *NAT1* star alleles (de León et al., 2000; Zhu et al., 2011; see also http://nat.mbg/duth.gr) when generating the GeT-RM consensus calls. For *CYP2E1* the low concordance was observed due to potentially novel star alleles that were not characterised by the GeT-RM study but were inferred by StellarPGx in ~40% of the participants in the Polaris HiSeqX cohort included for benchmarking analysis in this study. Some of the discrepancies in the star allele calls between StellarPGx and GeT-RM were supported by the results previously reported by Lee et al. (2019) using Stargazer. However, we found that StellarPGx had an added advantage in identifying participants with potentially novel star alleles.

Our analysis of SNVs across *CYP2E1, NAT1, NAT2, GSTM1*, and *GSTT1*, showed that SSA populations had a higher number of variants per kilobase compared to the other global populations included in this study. This is in keeping with the vast genetic diversity in Africa highlighted in other African-centred studies (Choudhury et al., 2017, 2020). Additionally, we identified a high number of potentially novel haplotypes (n=50) across the five pharmacogenes. This can be attributed to the fact that we analysed genomes from various underrepresented SSA populations and that previous studies focused on genotyping known star alleles. Based on the observations in our study and other SSA-based studies where novel haplotypes were identified in *CYP2E1* and *NAT2* (Patin et al, 2006; Heathfield et al., 2013), sampling other SSA populations may aid in obtaining a more comprehensive catalogue of star alleles in the pharmacogenes analysed in this study. Of the non-synonymous variants identified in these genes, 34% were predicted to be deleterious which warrants their inclusion in pharmacogenetic testing panels.

Among the *CYP2E1* variants, rs2070676 (present in the definition of **1B* and other haplotypes), rs2070673 and rs2070672 (defining **7C*) were observed at high frequencies (>15%) in the SSA populations. Additionally, of the non-synonymous variants used to define *CYP2E1* star alleles, rs6413419 (*CYP2E1*4*) was relatively African-specific. The high rs6413419 frequency observed in the various SSA populations (24.6%; ranging from 13.6 to 55%) and in the African American/Afro-Caribbean (22%) populations are similar to the frequency (20.6%) reported for African populations in gnomAD (Karczewski et al., 2020; Neyshaburinezhad et al., 2021). The rs6413419 variant is associated with normal activity (Fairbrother et al., 1998), and was predicted to be deleterious by only PolyPhen, CADD, LRT, and PROVEAN in this study. Many of the non-synonymous variants identified in the SSA populations were previously not used to define *CYP2E1* star alleles in literature. This made it difficult to resolve the majority of *CYP2E1* diplotypes in this study, leading to ambiguous calls in the WGS datasets harbouring potentially novel *CYP2E1* star alleles. The haplotypes harbouring these variants were predicted using StellarPGx and 14 were further characterised using trio data from the 1000G, where available. Six different *CYP2E1* haplotypes had rs28969387 as one of the defining variants. This variant was predicted to be deleterious in this study and has been predicted to alter the *CYP2E1* enzyme activity (Preissner et al., 2013). Given the incomprehensive *CYP2E1* star allele definitions, and the lack of more recent functional studies, it is clear that genotyping these SNVs among African individuals are important given the high frequency of some of these defining SNVs. Additionally, more studies are required in studying their impact on anti-TB drug response. With regard to *CYP2E1* structural variant analysis, SSA populations had a higher frequency of duplications compared to other global populations in this study. In particular we observed the **1C* duplication (**1Cx2*),

which has been previously described by Martis et al. (2013), in the SSA populations (AF=0.88%), but it was absent in the EUR and SAS populations analysed in this study. The duplication was reported at slightly higher frequencies in African American (3.8%), Hispanic (4.1%) and European-ancestry (2.9%) populations in the Martis et al. (2013) study. The *CYP2E1* gene deletion was observed in one individual from the MSL and the SEB population in this study. *CYP2E1* structural variants and their association with anti-TB drug response is understudied. However, they may be important to assay in pharmacogenetic tests for better optimisation of anti-TB drug dosing.

There is a scarceness of research regarding characterising *NAT1* star alleles, in the context of anti-TB pharmacogenomics, and the influence these star alleles may have on the response to anti-TB drugs such as PAS in both SSA and other global populations. The SSA populations had a significantly higher average frequency of *NAT1*10* (rapid acetylation allele*)*, compared to African American/Afro-Caribbean, AMR, SAS, and EUR. A previous study in South Africa consisting of predominantly Tswana-speaking individuals observed **10* at frequency of 50.5% (Loktionov et al., 2002), which is similar to the frequency across the SEB population (54.1%) in our study. Except for **4*, *NAT1* slow acetylation star alleles were less common compared to the rapid acetylation **10* allele in SSA. While this corresponds with previous studies in Africa (Loktionov et al., 2002; Sy et al., 2015), we observed relatively higher frequencies of key alleles such as *NAT1*3* (1.6-7.8%) given the diverse sampling in this study. For example, *NAT1*3*, which is associated with reduced clearance of PAS (Sy et al., 2015), has previously been observed in <1% of Black South African individuals (Loktionov et al., 2002; Sy et al., 2015). Despite the relatively low frequencies, **3* should be considered for pharmacogenetic testing given that it is associated with reduced clearance of PAS, a drug that is useful in treating individuals experiencing resistance to first-line anti-TB drugs. Sixteen possible novel *NAT1* haplotypes were identified in the SSA populations, of which 11 could be characterised further using trio data (Table 6, Supplementary Table 6). Regarding NAT1 predicted phenotype distributions, the acetylation profile in SSA differed from that observed in the EUR populations. The SSA populations had a higher proportion of individuals with a rapid acetylator phenotype while the slow acetylator phenotype was more frequent in EUR populations. This could be accounted for by variability in diet and environmental selection pressures (Mortensen et al., 2011).

Regarding *NAT2* variation, the SSA populations had a higher number of unique *NAT2* star alleles called in comparison to the other global populations in this study. In terms of the distribution between *NAT2* rapid and slow acetylator star alleles, slower acetylator star alleles, in particular **5B* (rs1801280, rs1799929, and rs1208) and **6A* (rs1041983 and rs1799930),

were more prevalent in both SSA and other global populations analysed in this study. This difference in the distribution of these two types of *NAT2* star alleles has been shaped by traditional subsistence, in which slow *NAT2* acetylators were positively selected for, and therefore increasing in frequency, as population groups transitioned from being hunter-gathers to pastoralists (Sabbagh et al., 2011). Given the high frequency of *NAT2* slow acetylation star alleles in SSA and other global population, the NAT2 slow acetylator phenotype was prevalent across the different population groups. However, the NAT2 predicted phenotype distribution in SSA still differed from the EAS populations. The NAT2 rapid acetylators contributed a small percentage to the phenotype profile in SSA, while NAT2 intermediate and slow acetylators were relatively equal being found in more than 40% of the population. In contrast, the three NAT2 acetylator phenotypes were seen in at least 30% of the individuals for each of the respective phenotypes. The differences in phenotypic profile between these two population groups may suggest different selective evolutionary events (Sabbagh et al., 2011).

The high prevalence of *GSTM1*0* (43.7-56.9%) and *GSTT1*0* (57.5-73.5%) in the various SSA study populations, is remarkable given the importance of GSTM1 and GSTT1 in detoxification of metabolites and toxins in the liver (Sanni et al., 2020). The high occurrence of these two gene deletions is comparable to other SSA-based population studies (Kassogue et al., 2020; Kasthurinaidu et al., 2015), however, there are differences in their distribution as noted in this study as well as others. Generally, *GSTM1*0* is less frequent compared to *GSTT1*0* in SSA. Notably, AMR, EAS, and EUR populations in this study had higher frequencies of *GSTM1*0* than the frequency in SSA, while the reverse was true for *GSTT1*0*. A similar pattern in the distribution of these alleles has been observed in other studies (Kasthurinaidu et al., 2015). A similar pattern was seen for the GSTM1 and GSTT1 null phenotypes, in which a higher proportion of individuals in SSA had the GSTT1 null phenotype while in the EUR population GSTM1 null phenotype was higher. This observation has been characterised in various studies, in which different African populations had a higher frequency of the GSTT1 null phenotype and GSTM1 present phenotype while the opposite was seen for some European populations (Palma-Cano et al., 2017; Piacentini et al., 2011). The differences in the distribution of these two deletions have been speculated to be due to differences in selection pressures, lifestyle habits, exposure to toxins and susceptibility to disease (Piacentini et al., 2011). The frequency of the *GSTM1-GSTT1* double homozygous deletion was low in various of the global populations, including in SSA (7.3%). There is limited known prevalence of the double *GSTM1-GSTT1* deletion across Sub-Saharan Africa compared to North African countries (Salem et al., 2011). In this study, the prevalence of the double null was lower in comparison to the relatively higher prevalence in North African countries (14.4-44%) (Salem et al., 2011; Chbili et al., 2022). Of the existing distribution among the various SSA populations

a higher prevalence has been observed in Burkina Faso (14.98%), Ivory Coast (14.3%) and Mali (11.2%) (Salem et al., 2011; Sombié et al., 2020; Kassogue et al., 2020). The double null genotype has been shown to be associated with AT-DILI in South American populations (Jaramillo-Valverde et al., 2022). To our knowledge a similar association has only been identified in Tunisia, a North African country (Chbili et al., 2022). Further studies on the influence of the *GSTM1-GSTT1* null deletion on anti-TB drug response in underrepresented SSA population groups is needed, given the relatively high frequency the deletion in SSA.

There were some limitations to this study. Firstly, our study population largely consisted of ethnolinguistic groups from the Niger-Congo language family and the sample sizes of the individual sub-groups were relatively small. Nonetheless, the findings from this study (e.g. novel star allele predictions) and the methodology (using StellarPGx and WGS data) lay the foundation for further star allele characterisations and functional studies in Africa relevant to anti-TB metabolism pharmacogenes. The use of larger datasets with more diverse underrepresented African populations would be critical and informative for designing effective pharmacogenetic testing strategies across the continent. Secondly, the lack of curated and comprehensive star allele definitions, especially in the case of *CYP2E1*, led to a number of ambiguous diplotype calls. This made it difficult to compare haplotypes called in this study with those previously published in literature. For *CYP2E1* we opted to focus on variant-level rather than haplotype-level analysis, while for *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* which had better curated star allele nomenclatures, StellarPGx was able to provide plausible star allele calls. Thirdly, this study used short read sequencing data, which made it difficult to phase variants defining possible novel haplotypes. This was resolved by using trio data where available. Furthermore, the fact that the genomic data was high depth (≥30x) provided reliable variant quality metrics. Fourthly, functional studies to confirm the clinical impact of the variants predicted to be deleterious by *in silico* prediction tools, was out of the scope of this study. Lastly, the phenotype predictions in this study were based solely on genetic variation and may differ from observed anti-TB drug response phenotypes due to other factors such as age, diet, environmental factors, and phenoconversion which could not be accounted for in this study.

Future directions

This study mainly focused on characterisation of the pharmacogenetic variation landscape across *NAT1*, *NAT2*, *GSTM1, GSTT1* and *CYP2E1* across African populations. The reported phenotypes for the enzymes encoded by these genes were based on predictions from the participant diplotypes. A vital next step would be to establish whether these predicted INH response phenotype patterns correspond to observations in clinical-based INH

pharmacogenomics studies involving the SSA populations represented in this study. This would aid in understanding the link of anti-TB drug response to the associated activity of the star alleles identified in this study.

Future work entailing the validation of novel haplotypes predicted in this study using other sequencing approaches (Sanger sequencing and/or long-read sequencing) is necessary to further characterise their sequence content and support their designation as novel alleles by PharmVar and aid in determining the possible impact that these novel haplotypes may have to INH response.

Conclusion

In conclusion, this study elucidates the pharmacogenomic variation in *CYP2E1*, *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* across various SSA populations, and in comparison, with other global populations. This study highlights differences of key star alleles (e.g. *NAT1*10*, *NAT2*5B*, *GSTM1*0*, and *GSTT1*0*) between SSA populations and other global populations. Furthermore, the high number of novel alleles, many being defined by functionally relevant variants, exemplify the need for more diverse sampling and large scale pharmacogenomic studies in Africa. Our findings on the predicted phenotype distributions warrant the need for further investigations on the potential for genotype-guided dosing regimens for INH and other anti-TB drugs in mitigating the risk of ADRs such as AT-DILI in SSA and global settings. In line with this, the information on the diversity of common, rare and novel star alleles among various SSA populations is critical in developing effective pharmacogenetic testing panels for use in Africa.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

The 1000 Genomes Project (www.internationalgenome.org/) and SGDP datasets (https://www.simonsfoundation.org/simons-genome-diversity-project/) are publicly available. The high coverage genomes from the SAHGP (n=15) (EGAD00001003791) and the Africa Wits-INDEPTH Partnership for Genomic Research study (n=159) (EGAD00001006418, EGAD00001004448) 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.

Author contributions

TVB Malinga, D Twesigomwe and H Othman designed the study. TVB Malinga, D Twesigomwe and H Othman contributed to acquisition and interpretation of the data. TVB Malinga analysed the data and drafted the manuscript. TVB Malinga, D Twesigomwe and H Othman revised the manuscript for important intellectual content. D Twesigomwe and H Othman supervised the research. All authors approved the final version of the manuscript.

Funding

This research is a substudy under the AGORA-TM project which is funded through the Genomic Research Approach for Diversity and Optimising Therapeutics (GRADIENT) initiative, a partnership between GlaxoSmithKline, Novartis, and the South African Medical Research Council.

References

Abera, W., Cheneke, W., & Abebe, G. (2016). Incidence of antituberculosis-drug-induced hepatotoxicity and associated risk factors among tuberculosis patients in Dawro Zone, South Ethiopia: A cohort study. *Int J Mycobacteriol*. 5, 14–20.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr Protoc Hum Genet.* Chapter 7, Unit7.20.

Azuma, J., Ohno, M., Kubota, R., Yokota, S., Nagai, T., Tsuyuguchi, K., et al. (2013). NAT2 genotype guided regimen reduces isoniazid-induced liver injury and early treatment failure in the 6-month four-drug standard treatment of tuberculosis: A randomized controlled trial for pharmacogenetics-based therapy. *Eur J Clin Pharmacol*. 69, 1091–1101.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.

Boyce, M. R., Katz, R., & Standley, C. J. (2019). Risk factors for infectious diseases in urban environments of sub-Saharan Africa: A systematic review and critical appraisal of evidence. In *Trop Med Infect Dis*. 4, 123.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 185, 3426–3440.

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*(Suppl 3), S3.

Cavaco, M. J., Alcobia, C., Oliveiros, B., Mesquita, L. A., Carvalho, A., Matos, F., et al. (2022). Clinical and Genetic Risk Factors for Drug-Induced Liver Injury Associated with Anti-Tuberculosis Treatment—A Study from Patients of Portuguese Health Centers. *J Pers Med.* 12, 790.

Chbili, C., Fathallah, N., Laadhari, C., Ouni, B., Saguem, S., Ben Fredj, M., et al. (2022). Glutathione-S-transferase genetic polymorphism and risk of hepatotoxicity to antitubercular drugs in a North-African population: A case-control study. *Gene*. 809, 146019.

Chen, R., Wang, J., Zhang, Y., Tang, S., & Zhan, S. (2015). Key factors of susceptibility to anti-tuberculosis drug-induced hepatotoxicity. *Arch Toxicol.* 89, 883–897.

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 31, 2745.

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., et al. High-depth African genomes inform human migration and health. *Nature*. 586, 741–748.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat Commun.* 8, 1–12.

Costa, G. N. O., Magno, L. A. V., Santana, C. V. N., Konstantinovas, C., Saito, S. T., Machado, M., et al. (2012). Genetic interaction between NAT2, GSTM1, GSTT1, CYP2E1, and environmental factors is associated with adverse reactions to anti-tuberculosis drugs. *Mol Diagn Ther.* 16, 241–250.

de León, J. H., Vatsis, K. P., & Weber, W. W. (2000). Characterization of naturally occurring and recombinant human N-acetyltransferase variants encoded by NAT1. *Mol Pharmacol. 58*, 288–299.

Ebeshi, B. U., Bolaji, O. O., & Masimirembwa, C. M. (2011). Arylamine N-acetyltransferase 2 (NAT2) single nucleotide polymorphisms' frequencies in Nigerian populations. *Afr J Pharm Pharmacol Res.* 1, 1–6.

El-Jaick, K. B., Ribeiro-Alves, M., Soares, M. V. G., Araujo, G. E. F. de, Pereira, G. R. C., Rolla, V. C., et al. (2022). Homozygotes NAT2*5B slow acetylators are highly associated with hepatotoxicity induced by anti-tuberculosis drugs. *Mem Inst Oswaldo Cruz.* 117, e210328.

Fairbrother, K. S., Grove, J., de Waziers, I., Steimel, D. T., Day, C. P., Crespi, C. L., & Daly, A. K. (1998). Detection and characterization of novel polymorphisms in the CYP2E1 gene. *Pharmacogenetics*. 8, 543–552.

Fujita, A., Kojima, K., Patriota, A. G., Sato, J. R., Severino, P., Miyano, S., & Dopazo, J. (2010). A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets. *Bioinformatics.* 26, 2349–2351.

Haas, D. W., Abdelwahab, M. T., van Beek, S. W., Baker, P., Maartens, G., Bradford, Y., et al. (2022). Pharmacogenetics of Between-Individual Variability in Plasma Clearance of Bedaquiline and Clofazimine in South Africa. *J Infect Dis.* 226, 147–156.

Hayashi, S.-I., Watanabe, J., & Kawajiri, K. (1991). Genetic Polymorphisms in the 5'-Flanking Region Change Transcriptional Regulation of the Human Cytochrome P450IIE1 Gene 1. *J Biochem*. 110, 559–565.

Heathfield, L. J., Dalvie, S., Kalideen, K., & Dandara, C. (2014). Novel CYP2E1 haplotype identified in a South African cohort. S. Afr. j. 110, 1–5.

Hu, Y., Hakkola, J., Oscarson, M., & Ingelman-Sundberg, M. (1999). Structural and functional characterization of the 5'-flanking region of the rat and human cytochrome P450 2E1 genes: identification of a polymorphic repeat in the human gene. *Biochem Biophys Res Commun*. *263*, 286–293.

Jaramillo-Valverde, L., Levano, K. S., Tarazona, D. D., Vasquez-Dominguez, A., Toledo-Nauto, A., Capristano, S., et al. (2022). GSTT1/GSTM1 Genotype and Anti-Tuberculosis Drug-Induced Hepatotoxicity in Peruvian Patients. *Int J Mol Sci*. 23, 11028.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 581, 434–443.

Kassogue, Y., Diakite, B., Kassogue, O., Konate, I., Tamboura, K., Diarra, Z., et al. (2020). Genetic polymorphism of drug metabolism enzymes (GSTM1, GSTT1 and GSTP1) in the healthy Malian population. *Mol Biol Rep*. 47, 393–400.

Kasthurinaidu, S. P., Ramasamy, T., Ayyavoo, J., Dave, D. K., & Adroja, D. A. (2015). GST M1-T1 null allele frequency patterns in geographically assorted human populations:A phylogenetic approach. *PLoS ONE.* 10, e0118660.

Kesenogile, B., Godman, B., & Rwegerera, G. M. (2021). Alanine transaminase and hemoglobin appear to predict the occurrence of antituberculosis medication hepatotoxicity; findings and implications in Botswana. *Expert Rev Anti Infect Ther*. 19, 379–391.

Lee, S. been, Wheeler, M. M., Thummel, K. E., & Nickerson, D. A. (2019). Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences. *Clin Pharmacol Ther*. 106, 1328–1337.

Levano, K. S., Jaramillo-Valverde, L., Tarazona, D. D., Sanchez, C., Capristano, S., Vásquez-Loarte, T., et al. (2021). Allelic and genotypic frequencies of NAT2, CYP2E1, and AADAC genes in a cohort of Peruvian tuberculosis patients. *Mol Genet Genomic Med*. 9, e1764.

Loktionov, A., Moore, W., Spencer, S. P., Vorster, H., Nell, T., O'neill, I. K., Bingham, S. A., & Cummings, J. H. (2002). Differences in N-acetylation genotypes between Caucasians and Black South Africans: implications for cancer prevention. *Cancer Detect Prev.* 26, 15–22.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 538, 201-206.

Martis, S., Mei, H., Vijzelaar, R., Edelmann, L., Desnick, R. J., & Scott, S. A. (2013). Multi-ethnic Cytochrome-P450 Copy Number Profiling: Novel Pharmacogenetic Alleles and Mechanism of Copy Number Variation Formation. *Pharmacogenomics J*. 13, 558-566.

Martis, S., Peter, I., Hulot, J.S., Kornreich R., Desnick, R.J., & Scott S.A. (2013). Multi-ethnic distribution of clinically relevant *CYP2C* genotypes and haplotypes. *Pharmacogenomics J*. 13, 369–377.

Mcbride, O. W., Umeno, M., Gelboin, H. V, & Gonzalez, F. J. (1987). A Taq I polymorphism in the human P450HE1 gene on chromosome 10 (CYP2E). *Nucleic Acids Res*. 15, 10071.

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

Mortensen, H. M., Froment, A., Lema, G., Bodo, J. M., Ibrahim, M., Nyambo, T. B et al. (2011). Characterization of genetic variation and natural selection at the arylamine N-acetyltransferase genes in global human populations. *Pharmacogenomics.* 12, 1545–1558.

Neyshaburinezhad, N., Ghasim, H., Rouini, M., Daali, Y., & Ardakani, Y. H. (2021). Frequency of important CYP450 enzyme gene polymorphisms in the Iranian population in comparison with other major populations: A comprehensive review of the human data. *J Pers Med*. 11, 804.

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

Palma-Cano, L. E., Córdova, E. J., Orozco, L., Martínez-Hernández, A., Cid, M., Leal-Berumen, I et al. (2017). GSTT1 and GSTM1 null variants in mestizo and amerindian populations from Northwestern Mexico and a literature review. *Genet Mol Biol*. 40, 727-735.

Patin, E., Harmant, C., Kidd, K. K., Kidd, J., Froment, A., Mehdi, S. Q., et al. (2006). Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene. *Human mutation*. *27*, 720.

Persson, I., Johansson, I., Bergung, H., Dahl, M. L., Seidegård, J., Rylander, R., et al. (1993). Genetic polymorphism of cytochrome P4502E1 in a Swedish population. Relationship to incidence of lung cancer. *FEBS Lett*. 319, 207–211.

Piacentini, S., Polimanti, R., Porreca, F., Martínez-Labarga, C., De Stefano, G. F., & Fuciarelli, M. (2011). GSTT1 and GSTM1 gene polymorphisms in European and African populations. *Mol Biol Rep*. 38, 1225–1230.

Pratt, V. M., Everts, R. E., Aggarwal, P., Beyer, B. N., Broeckel, U., Epstein-Baak, R., et al. (2016). Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. *J Mol Diagn*. 18, 109–123.

Preissner, S. C., Hoffmann, M. F., Preissner, R., Dunkel, M., Gewiess, A., & Preissner, S. (2013). Polymorphic cytochrome P450 enzymes (CYPs) and their role in personalized therapy. *PLoS ONE*. 8, eB2562.

Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., et al (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. *Glob Health Epidemiol Genom*. 1,e20.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 47, D886-D984.

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. 39, e118.

Rodrigues, C. H. M., Pires, D. E. V., & Ascher, D. B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci*. 30, 60–69.

Sabbagh, A., Darlu, P., Crouau-Roy, B., & Poloni, E. S. (2011). Arylamine N-Acetyltransferase 2 (NAT2) Genetic Diversity and Traditional Subsistence: A Worldwide Population Survey. *PLoS ONE.* 6, e18507.

Salem, A. H., Yaqoob, A., Ali, M., Handu, S., Fadel, R., Abu-Hijleh, M., & Almawi, W. (2011). Genetic polymorphism of the glutathione S-transferase M1 and T1 genes in three distinct Arab populations. *Dis Markers*. 31, 311–316.

Samarasinghe, S. R., Hoy, W., Jadhao, S., McMorran, B. J., Guchelaar, H. J., & Nagaraj, S. H. (2023). The pharmacogenomic landscape of an Indigenous Australian population. *Front Pharmacol.* 14, 1180640.

Sanni, S., Sina, H., & Baba-Moussa, L. (2022). Genetic Polymorphisms and Toxicities of First-Line Antituberculosis Drugs: Systematic Review of the Literature. Journal of Tuberculosis Research, 10, 124–145.

Sombié, H. K., Sorgho, A. P., Kologo, J. K., Ouattara, A. K., Yaméogo, S., Yonli, A. T., et al. (2020). Glutathione S-transferase M1 and T1 genes deletion polymorphisms and risk of developing essential hypertension: A case-control study in Burkina Faso population (West Africa). *BMC Med Genet*. 21, 55.

Sy, S. K. B., De Kock, L., Diacon, A. H., Werely, C. J., Xia, H., Rosenkranz, B., et al (2015). N-acetyltransferase genotypes and the pharmacokinetics and tolerability of para-aminosalicylic acid in patients with drug-resistant pulmonary tuberculosis *Antimicrob Agents Chemother*. 59, 4129–4138.

Tommaso, P. DI, Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 35, 316–319.

Toure, A., Cabral, M., Niang, A., Diop, C., Garat, A., Humbert, L., et al (2016). Prevention of isoniazid toxicity by NAT2 genotyping in Senegalese tuberculosis patients. *Toxicol Rep*. 3, 826-831.

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

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. *Clin Pharmacol Ther.* 110, 741–749.

Wang, D., Para, M. F., Koletar, S. L., & Sadee, W. (2011). Human N-acetyltransferase 1 (NAT1) *10 and *11 alleles increase protein expression via distinct mechanisms and associate with sulfamethoxazole-induced hypersensitivity. *Pharmacogenetics and genomics*. 21, 652-664. doi.org/10.1097/FPC.0B013E3283498EE9

World Health Organization (WHO). 2022. Global Tuberculosis report 2022. Available: http://apps.who.int/bookorders [Accessed: 11.02.2023].

Zenebe, Y., Adem, Y., Mekonnen, D., Derbie, A., Bereded, F., Bantie, M., et al. (2016). Profile of tuberculosis and its response to anti-TB drugs among tuberculosis patients treated under the TB control programme at Felege-Hiwot Referral Hospital, Ethiopia. *BMC Public Health.* 16, 688.

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 J*. 19, 115–126.

Zhu, Y., States, J. C., Wang, Y., & Hein, D. W. (2011). Functional effects of genetic polymorphisms in the N-acetyltransferase 1 coding and 3′ untranslated regions. *Birth Defects Res A Clin Mol Teratol*. 91, 77–84. doi.org/10.1002/bdra.20763

Supplementary Table 1: Diplotype calls made by StellarPGx and Stargazer in comparison to reference materials from the CDC's Genetic Testing Reference Materials Coordination Program (GeT-RM)

Corriell ID	*CYP2E1*			*NAT1*			*NAT2*			*GSTM1*			*GSTT1*		
	StellarPGx	GeT-RM	Stargazer	StellarPGx	GeT-RM	Stargazer	StellarPGx	GeT-RM	Stargazer	StellarPGx	GeT-RM	Stargazer	StellarPGx	GeT-RM	Stargazer
HG00276	*1A/*1A	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*5A/*6A	*5/*6	*5/*6	*0/*0	*0/*0	*0/*0	*0/*A	(*A/*A)	*A/*0
HG00436	[*5B/*5B]	(*7 / *7)	*7/*7	*10/*4	*4 / *4	*4/*10	*7B/*7B	*7/*7	*7/*7	*0/*0	*0/*0	*0/*0	*0/*0	*0/*0	*0/*0
HG00589	*5B/*7A	(*5)/ *7	*1/*7,*5	[*10/*10]	*4 / *4	*10/*26	*13A/*4	*4/(*13)	*4/*13	*0/*0	*0/*0	*0/*0	*0/*0	*0/*0	*0/*0
HG01190	*1A/[*1C] or *1B/*7B	*1 / *7	*1/*7	*10/*4	*4 / *4	*4/*10	*4/*4	*4/*4	*4/*4	*0/*0	*0/*0	*0/*0	*0/*0	*0/*0	*0/*0
NA06991	*1A/*1A	*1 / *1	*1/*1	[*10/*11B]	*4 / *11	*4/*11	*6A/*5B	*4/*5 or *5/*6	*5/*6	*A/*B	*B/*B	*A/*B	*0/*A	(*A/0)	*0/*A
NA07000	*1A/*1A	(*1/*1)	*1/*1	[*10/*11B]	(*4/*11)	*4/*11	*5B/*5B	(*5/*5)	*5/*5	*A/*A	(*A/*A)	*A/*A	*0/*A	(*A/*0)	*0/*A
NA07019	*1A/*1A	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*6A/*6A	*6/*6	*6/*6	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA07029	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*6A/*5B	*4/*5 or *5/*6	*5/*6	*0/*A	*A/*A	*0/*A	*0/*A	(*A/0)	*0/*A
NA07055	*1A/*1A	*1 / *1	*1/*1	*10/*17	*4 / *17	*10/*17	*5B/*5B	*5/*5	*5/*5	*0/*A	*A/*A	*0/*A	*A/*A	(*A/*A)	*A/*A
NA07056	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*6A/*6A	*6/*6	*6/*6	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA07348	*1A/[*1C] or *1B/*7C	*1 / *7	*1/*7	*4/*4	*4 / *4	*4/*4	*5B/*5B	*5/*5	*5/*5	*0/*A	*A/*A	*0/*A	*0/*A	(*A/*A)	*0/*A
NA07357	[*1C/*hap7]	(*7) / *7	*7/*7	*4/*4	*4 / *4	*4/*4	*5B/*6A	*4/*5	*5/*6	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA10831	*1A/[*1C] or *1B/*7B	*1 / *7	*1/*7	*4/*4	*4 / *4	*4/*4	*4/*5A	*4/*5	*4/*5	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA10851	*1A/*1A	*1 / *1	*1/*1	*10/*10	*4 / *4	*10/*10	*5B/*4	*4/*5	*4/*5	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA10854	*1A/*1A	*1 / *1	*1/*1	*17/*4	*4 / *17	*4/*17	*6A/*4	*4/*6	*4/*6	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA11839	*5B/*7A	(*5)/ *7	*1/*7,*5	*4/*4	*4 / *4	*4/*4	*5B/*6A	*4/*5	*5/*6	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*A/*0
NA18519	*1A/*1B	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*13A/*4	*4/(*13)	*4/*13	*A/*A	*A/*A	*A/*A	*0/*A	(*A/*AXN)	*A/*0
NA18524	[*1C/*hap7]	(*7) / *7	*7/*7	*10/*4	*4 / *4	*4/*10	*4/*4	*4/*4	*4/*4	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA19176	*1C/[*1C] or *7A/[*1C]	(*7) / *7	*7/*7	*10/*10	*4 / *4	*10/*10	*5B/*6A	*4/*5	*5/*6	*0/*A	*A/*A	*0/*A	*0/*0	0	*0/*0
NA12717	*1A/*1A	*1 / *1	*1/*1	*10/*3	*4 / *4	*3/*10	*4/*5B	*4/*5	*4/*5	*0/*A	*A/*A	*0/*A	*A/*A	(*A/*A)	*A/*A
NA19109	*1B/*1B	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*6A/*4	*4/*6	*4/*6	*A/*A	*A/*A	*A/*A	[*A/*A]	(*A/*AXN)	*A/*A
NA19174	[*1C/*1C]	(*7 / *7)	*7/*7	*10/*4	*4 / *4	*4/*10	*6A/*5B	*4/*5 or *5/*6	*5/*6	*0/*0	*0/*0	*0/*0	[*A/*A]	(*A/*AXN)	*A/*A
NA19178	*1C/[*1C]	(*7) / *7	*7/*7	*4/*4	*4 / *4	*4/*4	*5C/*6A	*5/*6	*5/*6	*A/*A	*A/*A	*A/*A	*0/*0	0	*0/*0
NA18509	[*1C/*1C]	(*7 / *7)	*7/*S1	*4/*4	*4 / *4	*4/*4	*12B/*13A	(*12/*13)	*12/*13	*0/*0	*0/*0	*0/*0	*0/*A	(*A/*A)	*0/*A

NA18518	*1A/*hap3 or *1B/*hap4	(*4) / *7	*1/*7, *4	*10/*4	*4 / *4	*4/*10	*14B/*4	(*4)/*14	*4/*14	*0/*A	*A/*A	*0/*A	*0/*A	(*A/*A)	*0/*A
NA18484	*1A/[*1C] or *1B/*7C	*1 / *7	*1/*7	*10/*10	*4 / *4	*10/*10	*4/*14A	*4/*14	*4/*14	*0/*A	*A/*A	*0/*A	*0/*A	(*A/*AXN)	*0/*A
NA19147	*1A/[*1C] or *1B/[*1C]	(*4) / *7	*1/*S1,*7,*4	*4/*4	*4 / *4	*4/*4	*5B/*13A	*4/*5	*5/*13	*A/*B	*A/*B	*A/*B	*0/*A	(*A/0)	*0/*A
NA18565	*1A/*1A	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*7B/*4	*4/*7	*4/*7	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA18959	*1A/*1A	*1 / *1	*1/*1	*10/*10	*4 / *4	*10/*10	*7B/*6A	*6/*7	*6/*7	*B/*B	*B/*B	*B/*B	*0/*A	(*A/*A)	*0/*A
NA18966	*1A/*1A	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*4/*4	*4/*4	*4/*4	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA18540	*5B/*7A	(*5)/ *7	*1/*7,*5	*4/*4	*4 / *4	*4/*4	*5B/*7B	*5/*7	*5/*7	*B/*B	*B/*B	*B/*B	*0/*0	0	*0/*0
NA19207	*1C/*1C	*7 / *7	*7/*7	*4/*4	*4 / *4	*4/*4	*14B/*12A	(*4)/*14	*12/*14	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA19143	[*1C/*1C]	(*7 / *7)	*7/*S1	*10/*4	*4 / *4	*4/*10	*13A/*14B	(*13)/*14	*13/*14	*0/*A	*A/*A	*0/*A	*0/*0	0	*0/*0
NA18855	*7C/[*1C] or [*1C/*1C]	(*7) / *7	*7/*7	[*10/*11A]	*4 / *11	*10/*11	*13A/*6A	*6/(*13)	*6/*13	*0/*A	*A/*A	*0/*A	[*A/*A]	(*A/*A)	*A/*A
NA12873	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5A/*5B	*5/*5	*5/*5	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA18992	[*1C/*1C]	(*7) / *7	*7/*7	*10/*4	*4 / *4	*4/*10	*4/*4	*4/*4	*4/*4	*0/*B	*B/*B	*0/*B	*0/*A	(*A/0)	*0/*A
NA18980	[*5B/*5B]	(*7 / *7)	*7/*7	*10/*4	*4 / *4	*4/*10	*4/*7B	*4/*7	*4/*7	*0/*A	*A/*A	*0/*A	*0/*A	(*A/*A)	*0/*A
NA11993	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5B/*5B	*5/*5	*5/*5	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA19239	*1C/[*1C] or *7A/[*1C]	(*7) / *7	*7/*7	*10/*4	*4 / *4	*4/*10	*12A/*14H	(*4)/*14	*12/*14	*A/*A	*A/*A	*A/*A	*0/*0	0	*0/*0
NA12003	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5B/*5B	*5/*5	*5/*5	*0/*0	*0/*0	*0/*0	*0/*A	(*A/*A)	*0/*A
NA18564	*5B/*7A	(*5)/ *7	*1/*7, *5	*4/*4	*4 / *4	*4/*4	*4/*7B	*4/*7	*4/*7	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA12156	*1A/*1A	*1 / *1	*1/*1	*10/*4	*4 / *4	*4/*10	*6A/*5D	*5/*6	*5/*6	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA12813	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5B/*6A	*4/*5 or *5/*6	*5/*6	*0/*0	*0/*0	*0/*0	*0/*A	(*A/*A)	*0/*A
NA18973	*1A/*7C	*1 / *7	*1/*7	*4/*4	*4 / *4	*4/*4	*4/*6A	*4/*6	*4/*6	*A/*B	*B/*B	*A/*B	*0/*A	(*A/*A)	*0/*A
NA18942	*5B/*7A	(*5)/ *7	*1/*7, *5	*10/*10	*4 / *4	*10/*10	*4/*6A	*4/*6	*4/*6	*0/*B	*B/*B	*0/*B	*0/*0	0	*0/*0
NA19789	*1A/*1A	*1 / *1	*1/*1	*10/*10	*4 / *4	*10/*10	*4/*4	*4/*4	*4/*4	*0/*B	*B/*B	*0/*B	*0/*A	(*A/0)	*0/*A
NA10847	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5A/*5B	*5/*5	*5/*5	*0/*A	*A/*A	*0/*A	*B/*0	(*B/0)	*B/*0
NA12006	*1A/*1A	*1 / *1	*1/*1	[*10/*11A]	*4 / *11	*10/*11	*6A/*6A	*6/*6	*6/*6	*0/*A	*A/*A	*0/*A	*A/*A	(*A/*A)	*A/*A
NA18526	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*4/*4	*4/*4	*4/*4	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA12145	*1A/*1A	*1 / *1	*1/*1	*4/*14A	*4 / *14	*4/*14	*4/*5B	*4/*5	*4/*5	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA18868	*1C/[*1C]	(*7) / *7	*7/*7	*10/*4	*4 / *4	*4/*10	*5C/*13A	*5/(*12)	*5,*12/*13	*A/*A	*A/*A	*A/*A	[*0/*A]	(*A/*A)	*A/*0
NA19095	[*1C]/*1Bx2	*1 / *7	*1/*7x2	*4/*4	*4 / *4	*4/*4	*6A/*6A	*6/*6	*6/*6	*A/*A	*A/*A	*A/*A	[*0/*A]	(*A/*AXN)	*A/*0

NA19122	*1C/[*1C]	*7 / *7	*7/*7	*10/*4	*4 / *4	*4/*10	*12A/*12H	(*12/*12)	*12/*12	*0/*A	*A/*A	*0/*A	*0/*A	(*A/*A)	*0/*A
NA19003	*1A/[*1C] or *1B/*7C	*1 / *7	*1/*7	*10/*4	*4 / *4	*4/*10	*4/*4	*4/*4	*4/*4	*0/*0	*0/*0	*0/*0	*0/*A	(*A/0)	*0/*A
NA19226	*1B/*1Cx2	*1 / *7	*1/*7x2	*4/*4	*4 / *4	*4/*4	*5B/*4	*4/*5	*4/*5	*A/*A	*A/*A	*A/*A	*0/*0	0	*0/*0
NA18544	[*1C/*1C]	*7 / *7	*7/*7	*10/*4	*4 / *4	*4/*10	*6A/*4	*4/*6	*4/*6	*0/*B	*B/*B	*0/*B	*0/*A	(*A/0)	*0/*A
NA18952	*1A/*1A	*1 / *1	*1/*1	*10/*10	*4 / *4	*10/*10	*4/*4	*4/*4	*4/*4	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA11832	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5B/*12A	*5/(*12)	*5/*12	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA18861	*1C/[*1C] or *7A/[*1C]	(*7) / *7	*7/*7	*10/*10	*4 / *4	*10/*10	*5B/*5B	*5/*5	*5/*5	*A/*B	*A/*B	*A/*B	*0/*A	(*A/*AXN)	*0/*A
NA19908	*7A/*1Cx3	*7 / *7	*7x2/*7x2	*4/*4	*4 / *4	*4/*4	*5B/*5C	*5/*5	*5/*5	[*A/*A]	*A/*A	*A/*Ax2	*0/*A	(*A/*A)	*0/*A
NA19819	*1A/*1C	*1 / *7	*1/*7	*10/*4	*4 / *4	*4/*10	*5C/*6O	*5/*6	*5/*6	*A/*B	*A/*B	*A/*B	*A/*A	(*A/*A)	*A/*A
NA19920	*1A/[*1C] or *1B/[*1C]	(*4) / *7	*1/*S1, *7, *4	*10/*4	*4 / *4	*4/*10	*6A/*6A	*6/*6	*6/*6	*0/*A	*A/*A	*0/*A	*0/*A	(*A/*AXN)	*0/*A
NA19007	*5B/*7A	(*5)/ *7	*1/*7, *5	*10/*10	*4 / *4	*10/*10	*7B/*4	*4/*7	*4/*7	*0/*0	*0/*0	*0/*0	*0/*0	0	*0/*0
NA19213	*7C/[*1C] or [*1C]/[*1C]	(*7) / *7	*7/*7	*10/*10	*4 / *4	*10/*10	*12A/*14H	(*4)/*14	*12/*14	*0/*A	*A/*A	*0/*A	*A/*A	(*A/*AXN)	*A/*A
NA18552	*5B/[*1C]	(*5)/ *7	*5/*7	*10/*10	*4 / *4	*10/*10	*4/*4	*4/*4	*4/*4	*0/*B	*B/*B	*0/*B	*0/*A	(*A/*A)	*0/*A
NA18617	[*5B/*5B]	*7/*7	*7/*7	*10/*10	(*4/*4)	*10/*10	*4/*6A	(*4/*6)	*4/*6	*0/*0	(*0/*0)	*0/*0	*0/*A	(*A/*A)	*0/*A
NA19917	*1B/*1C	(*1/*7)	*1/*7	*10/*4	(*4/*4)	*4/*10	*6H/*4	(*4/*6)	*4/*6	*A/*A	(*A/*A)	*A/*A	*0/*A	(*A/*A)	*0/*A
NA20296	*1A/*1B	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*10	*6A/*60	*6/*6	*6/*6	*0/*A	*A/*A	*0/*A	*A/*A	(*A/*A)	*A/*A
NA20509	*1A/*1A	*1 / *1	*1/*1	*4/*4	*4 / *4	*4/*4	*5B/*6A	*4/*5	*5/*6	*0/*0	*0/*0	*0/*0	*A/*A	(*A/*A)	*A/*A
NA21781	*1A/*1A	(*1/*1)	*1/*1	*4/*4	(*4/*4)	*4/*4	*5B/*6A		*5/*6	*0/*0	(*0/*0)	*0/*0	*0/*0	(*0/*0)	*0/*0

Square brackets indicate possible novel haplotypes flagged by StellarPGx;
Curved brackets indicate uncertain haplotype/diplotype calls

Supplementary Table 2: *In-silico* predictions made for prioritised allele defining variants for known and possible novel star alleles

Variant ID	Consequence	SIFT	PolyPhen	CADD	LRT	Mutation Assessor	PROVEAN	VEST4	Consensus
CYP2E1									
rs367957731	Missense (L17F)		✓	✓					
g.5167T>C	Missense (L45P)	✓	✓	✓		✓	✓	✓	✓
rs6413419	Missense (V179I) (**4*)		✓	✓	✓		✓		✓
rs749979006	Missense (R282H)	✓							
rs1007026512	Missense (D295N)		✓	✓	✓		✓		✓
rs759707874	Missense (R337Q)		✓	✓	✓				
rs59981143	Missense (T373A)			✓			✓		
rs59656378	Missense (V396L)								
rs57702102	Missense (N418S)								
rs142767992	Missense (S424G)	✓	✓	✓	✓	✓	✓		✓
rs199855848	Missense (M445I)	✓	✓	✓	✓	✓	✓	✓	✓
rs28969387	Missense (H457L)	✓	✓	✓		✓	✓		✓
rs555344732	Stop gain (R337*)			✓	✓			✓	✓
rs3813867	2KB Upstream (**5A*,**5B*)								
rs2031920	2KB Upstream (**5A*,**5B*)								
rs2070672	2KB Upstream (**7C*)								
rs2070673	2KB Upstream (**7C*)								
rs6413420	2KB Upstream (**7B*)								
rs2070676	Intron (**1B, *7A, *7B, *7C*)								
NAT1									
rs111848753	Missense (R33Q)								
rs138061602	Missense (R64Q)	✓	✓	✓	✓	✓	✓		✓
rs370859724	Missense (M105I)								
rs548004925	Missense (Y129C)				✓				

rs529845573	Missense (P147S)	✓	✓		✓	✓	✓		✓
rs4987076	Missense (V149I) (**11A*)								
rs370302501	Missense (R151H)		✓		✓	✓	✓		✓
rs554793519	Missense (L161P)	✓	✓	✓		✓	✓	✓	✓
rs4986783	Missense (S214A) (**11A*)								
rs201916736	Missense (V231F)				✓				
rs1389255907	Missense (G284D)			✓		✓			
g.57013~C>T	Stop gain (Q143*)			✓	✓			✓	✓
rs5030839	Stop gain (R187*) (**15*)			✓	✓				✓
rs4986992	Synonymous (L7L) (**27*)								
rs4986991	Synonymous (S259S) (**27*)								
rs4986990	Synonymous (T153T) (**11A*)								
rs146727732	Synonymous (P134P) (**20*)								
NAT2									
rs200893121	Missense (I3V) (**14L*)				✓				
rs45477599	Missense (L24I) (**5L*)	✓	✓	✓	✓			✓	✓
rs149283608	Missense (N41Y)(**12N*)	✓		✓	✓		✓		✓
rs1801279	Missense (R64Q) (**7D, *14A,*14B,*14D, *14E, *14G, *14H, *14L*)	✓	✓	✓	✓		✓		✓
rs561124342	Missense (G83A)	✓	✓	✓	✓		✓	✓	✓
rs1801280	Missense (I114T) (**5A, *5B, *5C, *5D, *5E, *5L, *5N, *5P, *5W*)	✓	✓				✓		
rs146405047	Missense (N118K)								
rs12720065	Missense (L135V) (**6P, *12H, *24*)								
rs139351995	Missense (I158L) (**5N*)								

rs79050330 (T193M)	Missense (T193M) (*5P, *12E*)	✓	✓	✓	✓		✓		✓
rs1799930	Missense (R197Q) (*5E, *5P, *6A, *6B, *6C, *6H, *6K, *6L,*6O, *6P, *7D,*14D*)		✓	✓	✓				
rs45618543	Missense (E203D) (*12G, *12O*)								
rs56387565	Missense (Y208H) (*12F*)		✓						
rs138707146	Missense (P213L) (*6K*)	✓	✓	✓	✓		✓	✓	✓
rs45518335	Missense (P228L) (*14H*)						✓		
rs55700793	Missense (K256E) (*6H*)		✓	✓					
rs1208	Missense (K268R) *(*5B, *5C, *5L, *5N, *5P, *5W, *6C, *12A, *12B, *12C, *12E, *12F, *12G, *12H, *12N, *12O, *14E, *14G*)								
rs868725509	Missense (I270T) (*26*)		✓		✓				
rs56393504	Missense (V280M) (*6O, *13F*)	✓							
rs1799931	Missense (G286E) (*7B, *7D*)								
rs375746304	Stop gain (R197*) (*27*)			✓	✓			✓	✓
rs1041983	Synonymous (Y94Y) (*5P, *12B, *13A, *6A, *6C, *6H, *6K, *6L, *7B, *14B, *14D, *14G*)								
rs45532639	Synonymous (D115D) (*6L*)								
rs146405047	Synonymous(N118N) (*5W*)								

rs1799929	Synonymous (L161L) *(*12C, *5A, *5B, *5L, *5N, *5P, *5W)*								
rs144176822	Synonymous (T193T) (**6U*)								
GSTM1									
rs184653774	Missense (D9E)	✓			✓	✓		✓	✓
rs146668816	Missense (N85K)				✓		✓		
rs138440339	Missense (L159F)					✓			
rs375154572	Missense (R168C)		✓	✓		✓	✓		✓
rs1065411	Missense (K173N)	✓				✓			
rs449856	Missense (S210T)								
rs533860247	Missense (A213T)	✓	✓	✓		✓			✓

The presence of a ✓ indicates the variant was predicted to be deleterious as it fit the following ADME-optimised criteria: SIFT < 0.0376, PolyPhen > 0.3841, CADD >19.19, LRT<0.0025, Mutation Assessor > 2.0566, PROVEAN < -3.286, VEST > 0.4534

Supplementary Table 3: Examples of trio data used to infer core *CYP2E1* (NG_008383.1) variants defining novel haplotypes

Coriell ID	Haplotypes	Allele count
HG02642 (Father)	Hap 1(**1C*+**4682~A>G**): 790~G>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;8663~T>C;8845~T>C;11061~C>A;11667~A>G;11861~A>G;12391~C>T;14259~GT>G;15496~C>T;15756~A>AT;17488~G>GAGGAGAATGGTGCCCAAGA;18728~C>T;19897~C>G;20385~T>TTC;20391~C>CTT	
	Hap 2 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;8883~A>C;11667~A>G;12506~AT>A;13515~G>T;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20288~CCTTTCTTT>C;20373~TTCTTTCTTTCTTTCTTTC>T	
HG02643 (Mother)	Hap 1 (**1B*+**14850~A>G**): 364~C>A;648~A>G;790~G>A;850~T>A;3519~T>G;4622~T>C;**4701~A>T**;5000~T>TC;6394~C>T;7872~C>T;8663~T>C;11667~A>G;12506~A>AT;13417~A>G;14259~GT>G;**14850~A>G**;15111~C>G;15496~C>T;15756~A>AT;15923~A>G;20385~T>TTC;20391~C>CTT	2
	Hap 2 (**1C*+**3739~G>C**+**9809~G>A**+**15271~G>C**): 790~G>A;1855~A>C;3468~T>A;**3739~G>C**;4007~T>C;4227~T>C;5000~T>TC;6481~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;11667~A>G;12506~AT>A;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14466~G>T;14499~C>T;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16648~A>G;16654~T>C;16989~T>C;17213~A>C;17633~T>G;18259~T>G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20298~TTTCTTTCTTTCTTTCTTTCTTTTCTTTCTTTC>T;20310~TTTCTTTCTTTTC>*;20366~CT>C;20389~T>C;20391~C>T;20435~T>C;20717~T>C;20928~G>T;21345~A>C;21491~C>T;21618~C>T;21757~C>T	
HG02644 (Child)	Hap 1 (**1C*+**4682~A>G**): 790~G>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;8663~T>C;8845~T>C;11061~C>A;11667~A>G;11861~A>G;12391~C>T;14259~GT>G;15496~C>T;15756~A>AT;17488~G>GAGGAGAATGGTGCCCAAGA;18728~C>T;19897~C>G;20385~T>TTC;20391~C>CTT	
	Hap 2 (**1B*+**14850~A>G**): 364~C>A;648~A>G;790~G>A;850~T>A;3519~T>G;4622~T>C;**4701~A>T**;5000~T>TC;6394~C>T;7872~C>T;8663~T>C;11667~A>G;12506~A>AT;13417~A>G;14259~GT>G;**14850~A>G**;15111~C>G;15496~C>T;15756~A>AT;15923~A>G;20385~T>TTC;20391~C>CTT	
Unphased variants: 6778~G>A;10830~C>T;12007~C>CTCG;12009~T>TGTC;12009~T>C		

HG02666 (Father)	Hap 1 (*1C+**3739~G>C**+**9809~G>A**+**15271~G>C**): 790~G>A;1855~A>C;**3739~G>C**;4007~T>C;5000~T>TC;6778~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;9886~T>C;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14466~G>T;14499~C>T;**15271~G>C**;15308~T>A;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16654~T>C;16989~T>C;17213~A>C;17633~T>G;18749~TAAGAA>T;19141~C>T;19415~C>CTCTTAAT;19416~A>G;20226~G>A;20297~C>CT;20389~T>C;20391~C>T;20435~T>C;20928~G>T;21345~A>C	
	Hap 2 (*1C+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;14260~T>TGGG;16825~A>T;16831~G>T;20310~TTTCTTTCTTTTCTTTCTTTC>T;20317~CT>*;20381~T>TTC;20385~TTC>T;20389~T>C;20391~C>T	
HG02667 (Mother)	Hap 1 (*1B*): 790~G>A;4622~T>C;**4701~A>T**;5000~T>TC;6394~C>T;7872~C>T;8663~T>C;11667~A>G;12009~T>TGTC;12506~A>AT;12506~AT>A;13417~A>G;15756~A>AT;15923~A>G;20288~CCTTTCTTT>C	
	Hap 2 (*1C+**16490~A>T**): 790~G>A;3642~T>G;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12007~C>CTCG;12009~T>C;14777~C>G;15120~C>T;15165~C>G;15756~AT>A;16220~G>T;16402~C>G;**16490~A>T**;16624~C>T;16734~A>C;16818~T>G;16825~A>G;16831~G>T;16832~T>G;16855~T>C;17112~C>T;17194~C>T;17344~G>A;18109~G>C;18214~T>C;18399~G>T;18728~C>T;19290~C>T;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19508~G>A;19518~C>G;19589~G>T;19783~A>G;19995~C>T;20385~T>TTC;20391~C>CTT;20728~G>T;20753~G>A;20818~T>C	8
HG02668 (Child)	Hap 1 (*1C+**3739~G>C** +**9809~G>A** +**15271~G>C**): 790~G>A;1855~A>C;**3739~G>C**;4007~T>C;5000~T>TC;6778~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;9886~T>C;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14466~G>T;14499~C>T;**15271~G>C**;15308~T>A;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16654~T>C;16989~T>C;17213~A>C;17633~T>G;18749~TAAGAA>T;19141~C>T;19415~C>CTCTTAAT;19416~A>G;20226~G>A;20297~C>CT;20389~T>C;20391~C>T;20435~T>C;20928~G>T;21345~A>C	
	Hap 2 (*1B*): 790~G>A;4622~T>C;**4701~A>T**;5000~T>TC;6394~C>T;7872~C>T;8663~T>C;11667~A>G;12009~T>TGTC;12506~A>AT;12506~AT>A;13417~A>G;15756~A>AT;15923~A>G;20288~CCTTTCTTT>C	

Unphased variants: 364~C>A;648~A>G;850~T>A;3468~T>A;3519~T>G;4227~T>C;15005~G>T;15496~C>T;15740~GTGT>G;16648~A>G;18259~T>G;18385~T>G;1884 8~G>C;19008~A>C;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20717~T>C;21491~C>T;21618~C>T;21757~C>T		
HG02678 (Father)	Hap 1 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;866 3~T>C;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;1662 4~C>T;16855~T>C;17112~C>T;18728~C>T;20288~CCTTTCTTT>C;20366~CT>C;20386~TC>T;20389~T>C;20391~C>T	
	Hap 2 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;866 3~T>C;8883~A>C;11667~A>G;12506~AT>A;14260~T>TGGG;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T> C;17112~C>T;18728~C>T;20288~CCTTTCTTT>C;20362~CTTTCT>C;20373~TTCTTTCTTTCTTTCTTTC>T;20380~TTTC>* ;20389~TTC>*	
HG02679 (Mother)	Hap 1 (**1C*+**10416~T>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8223~C>T;8663~T>C;**10416~T>A**;11667~A>G;12844~C>T;15120~C>T;15496 ~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20366~CT>C;20386~TC>T;20389~T>C;20391~C>T	1
	Hap 2 (**1A*): 790~G>A;2100~C>T;4672~C>T;**4701~A>T**;5000~T>TC;6394~C>T;6449~G>C;7872~C>T;8663~T>C;11667~A>G;12009~T> TGTC;12321~T>C;12506~AT>A;12891~C>G;13360~C>A;13671~G>A;13935~T>C;14048~C>T;14260~T>TGGG;14466~G> T;14499~C>T;14624~A>AGATGGGTGGAT;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G ;16234~T>C;16287~A>G;16302~T>C;16643~T>A;16648~A>G;16989~T>C;17003~T>C;17213~A>C;17633~T>G;18259~T> G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19 615~T>C;19668~A>G;19691~C>T;20310~TTTCTTTCTTTTCTTTC>T;20314~TTTCTTTTCTTTC>*;20385~T>TTC;20391~C >CTT;21373~A>C	
HG02680 (Child)	Hap 1 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;866 3~T>C;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;1662 4~C>T;16855~T>C;17112~C>T;18728~C>T;20288~CCTTTCTTT>C;20366~CT>C;20386~TC>T;20389~T>C;20391~C>T	
	Hap 2 (**1C*+**10416~T>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8223~C>T;8663~T>C;**10416~T>A**;11667~A>G;12844~C>T;15120~C>T;15496 ~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20366~CT>C;20386~TC>T;20389~T>C;20391~C>T	
	Hap 1 (**1C*+**4963~G>T**): 790~G>A;851~A>T;2530~T>G;**4963~G>T**;5000~T>TC;6069~C>T;6394~C>T;6778~G>A;8663~T>C;8713~A>G;9945~T>C;1	

HG02461 (Father)	0839~A>G;11667~A>G;14260~T>TGGG;15496~C>T;15756~A>AT;18728~C>T;20288~CCTTT>C;20366~CT>C;20376~TTTC>T;20389~TTC>T;20444~GTTTC>G	
	Hap 2 (*1C*): 790~G>A;3642~T>G;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20309~C>CT;20318~T>TTTC;20380~TTTC>T;20389~TTC>T	
HG02462 (Mother)	Hap 1 (*7A*+**3739~G>C**): 790~G>A;**3739~G>C**;5000~T>TC;6394~C>T;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;12391~C>T;12506~AT>A;12678~T>A;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13501~G>A;13671~G>A;13701~G>A;13935~T>C;14048~C>T;14260~T>TGGG;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;15756~AT>A;19337~C>T;20366~CT>C;20376~TTTC>T;20389~TTC>T;21067~G>A	2
	Hap 2 (*1A*): 790~G>A;2100~C>T;**4701~A>T**;5000~T>TC;6394~C>T;6449~G>C;7872~C>T;8663~T>C;11667~A>G;12321~T>C;12891~C>G;13360~C>A;13671~G>A;13935~T>C;14048~C>T;14466~G>T;14624~A>AGATGGGTGGAT;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16643~T>A;16648~A>G;16989~T>C;17003~T>C;17213~A>C;17633~T>G;18259~T>G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19615~T>C;19668~A>G;19691~C>T;20313~CTTTCT>C;20385~T>TTC;20391~C>CTT;20448~C>G;21373~A>C	
HG02463 (Child)	Hap 1 (*1C*+**4963~G>T**): 790~G>A;851~A>T;2530~T>G;**4963~G>T**;5000~T>TC;6069~C>T;6394~C>T;6778~G>A;8663~T>C;8713~A>G;9945~T>C;10839~A>G;11667~A>G;14260~T>TGGG;15496~C>T;15756~A>AT;18728~C>T;20288~CCTTT>C;20366~CT>C;20376~TTTC>T;20389~TTC>T;20444~GTTTC>G	
	Hap 2 (*7A*+**3739~G>C**) :790~G>A;**3739~G>C**;5000~T>TC;6394~C>T;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;12391~C>T;12506~AT>A;12678~T>A;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13501~G>A;13671~G>A;13701~G>A;13935~T>C;14048~C>T;14260~T>TGGG;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;15756~AT>A;19337~C>T;20366~CT>C;20376~TTTC>T;20389~TTC>T;21067~G>A	
Unphased variants: 364~C>A;648~A>G;850~T>A;3468~T>A;3519~T>G;4227~T>C;12007~C>CTCG;12009~T>TGTC;12009~T>C		
HG02464 (Father)	Hap 1 (*1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;14260~T>TGGG;15496~C>T;20389~T>C	
	Hap 2 (*1B*): 790~G>A;1304~G>C;4622~T>C;**4701~A>T**;5000~T>TC;6787~G>A;7872~C>T;8663~T>C;11667~A>G;12009~T>TGTC;125	

	06~A>AT;12844~C>T;13417~A>G;15496~C>T;15756~A>AT;15923~A>G;20309~C>CT;20318~T>TTTC;20381~T>TTC;20385~TTC>T;20389~T>C;20391~C>T	
HG02465 (Mother)	Hap 1 (**7C*+**5082~C>T** + **16490~A>T**): 364~C>A;648~A>G;790~G>A;850~T>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;3519~T>G;4104~A>G;**4682~A>G**;5000~T>TC;**5082~C>T**;6778~G>A;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>C;12391~C>T;12678~T>A;12713~A>G;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14260~T>TGGG;15005~G>T;15165~C>G;**15271~G>C**;15308~T>A;15740~GTGT>G;15756~AT>A;16220~G>T;16402~C>G;**16490~A>T**;16648~A>G;16734~A>C;16818~T>G;16825~A>G;16831~G>T;16832~T>G;17194~C>T;17344~G>A;18109~G>C;18202~G>A;18214~T>C;18385~T>G;18399~G>T;18848~G>C;19008~A>C;19290~C>T;19302~C>G;19337~C>T;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19760~TAA>T;19783~A>G;19995~C>T;20818~T>C;21067~G>A;21618~C>T;21757~C>T	1
	Hap 2 (**1C*+**4682~A>G**): 364~C>A;648~A>G;790~G>A;850~T>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;3519~T>G;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;6778~G>A;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>C;12678~T>A;12713~A>G;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14162~G>A;14213~A>G;14260~T>TGGG;14363~ATGGTTGGATGGATGGGTGGGTGGGTGGATGGATGGG>A;14499~C>T;15005~G>T;15740~GTGT>G;15756~AT>A;18259~T>G;20226~G>A;20381~T>TTC;20389~TTC>T;20390~TC>*;21491~C>T	
HG02466 (Child)	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;14260~T>TGGG;15496~C>T;20389~T>C	
	Hap 2 (**7C*+**5082~C>T**+**16490~A>T**): 364~C>A;648~A>G;790~G>A;850~T>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;3519~T>G;4104~A>G;**4682~A>G**;5000~T>TC;**5082~C>T**;6778~G>A;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>C;12391~C>T;12678~T>A;12713~A>G;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14260~T>TGGG;15005~G>T;15165~C>G;**15271~G>C**;15308~T>A;15740~GTGT>G;15756~AT>A;16220~G>T;16402~C>G;**16490~A>T**;16648~A>G;16734~A>C;16818~T>G;16825~A>G;16831~G>T;16832~T>G;17194~C>T;17344~G>A;18109~G>C;18202~G>A;18214~T>C;18385~T>G;18399~G>T;18848~G>C;19008~A>C;19290~C>T;19302~C>G;19337~C>T;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19760~TAA>T;19783~A>G;19995~C>T;20818~T>C;21067~G>A;21618~C>T;21757~C>T	
	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;	

HG02804 (Father)	9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;20385~T>C;20387~C>T;20389~T>C;20391~C>T	
	Hap 2 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;20385~T>C;20387~C>T;20389~T>C;20391~C>T	
HG02805 (Mother)	Hap 1 (**1B*+**16490~A>T**): 790~G>A;1304~G>C;4622~T>C;**4701~A>T**;5000~T>TC;6787~G>A;7872~C>T;8663~T>C;10830~C>T;11667~A>G;12009~T>TGTC;12506~A>AT;12844~C>T;13417~A>G;13773~G>A;14777~C>G;15005~G>T;15165~C>G;15496~C>T;15756~A>AT;15923~A>G;16402~C>G;**16490~A>T**;16734~A>C;16818~T>G;16825~A>G;16831~G>T;16832~T>G;17194~C>T;18109~G>C;18202~G>A;18214~T>C;18399~G>T;18848~G>C;19008~A>C;19290~C>T;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19518~C>G;19528~G>A;19534~C>T;19545~C>A;19589~G>T;19760~TA>T;19783~A>G;19995~C>T;20226~G>A;20305~C>CT;20314~T>C;20317~CT>C;20385~T>C;20387~C>T;20389~T>C;20391~C>T;20717~T>C;20728~G>T;20818~T>C;21491~C>T;21618~C>T;21757~C>T	3
	Hap 2 (**7C*): 364~C>A;648~A>G;790~G>A;850~T>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;3519~T>G;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;6717~C>T;6778~G>A;8663~T>C;8845~T>C;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>C;12160~G>A;12391~C>T;12678~T>A;12713~A>G;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13466~C>T;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14260~T>TGGG;14499~C>T;15271~G>C;15308~T>A;15740~GTGT>G;15756~AT>A;16220~G>T;16648~A>G;17344~G>A;18255~C>T;19337~C>T;20061~G>A;20288~CCTTTCTTTCTTT>C;20362~CTTTCT>C;20388~T>C;20390~TC>T;20435~T>C;20753~G>A	
HG02806 (Child)	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;20385~T>C;20387~C>T;20389~T>C;20391~C>T	
	Hap 2 (**1B*+**16490~A>T**): 790~G>A;1304~G>C;4622~T>C;**4701~A>T**;5000~T>TC;6787~G>A;7872~C>T;8663~T>C;10830~C>T;11667~A>G;12009~T>TGTC;12506~A>AT;12844~C>T;13417~A>G;13773~G>A;14777~C>G;15005~G>T;15165~C>G;15496~C>T;15756~A>AT;15923~A>G;16402~C>G;**16490~A>T**;16734~A>C;16818~T>G;16825~A>G;16831~G>T;16832~T>G;17194~C>T;18109~G>C;18202~G>A;18214~T>C;18399~G>T;18848~G>C;19008~A>C;19290~C>T;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19518~C>G;19528~G>A;19534~C>T;19545~C>A;19589~G>T;19760~TA>T;19783~A>G;19995~C>T;20226~G>A;20305~C>CT;20314~T>C;20317~CT>C;20385~T>C;20387~C>T;20389~T>C;20391~C>T;20717~T>C;20728~G>T;20818~T>C;21491~C>T;21618~C>T;21757~C>T	

Unphased variant: 10152~A>G		
HG02570 (Father)	Hap 1 (*1C+**9809~G>A**): 364~C>A;648~A>G;734~ATT>A;790~G>A;850~T>A;851~A>T;3519~T>G;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10839~A>G;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12506~AT>A;15496~C>T;15756~A>AT;18728~C>T	
	Hap 2 (*1C+**3739~G>C**+**9809~G>A**+**15271~G>C**): 790~G>A;1855~A>C;**3739~G>C**;4007~T>C;5000~T>TC;6778~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;9886~T>C;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12506~AT>A;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14260~T>TGGG;14466~G>T;14499~C>T;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16648~A>G;16654~T>C;16989~T>C;17213~A>C;17633~T>G;18259~T>G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20226~G>A;20301~C>T;20302~T>C;20305~C>T;20306~T>C;20309~C>T;20310~T>C;20313~C>T;20314~T>C;20317~C>T;20318~T>C;20377~TTCTTTCTTTC>T;20389~T>C;20391~C>T;20435~T>C;20717~T>C;20928~G>T;21345~A>C;21491~C>T;21618~C>T;21757~C>T	
HG02571 (Mother)	Hap 1(*1C): 790~G>A;3642~T>G;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20373~TTCTTTCTTTCTTTCTTTC>T	
	Hap 2 (*1C+**4682~A>G**+**16490~A>T**): 364~C>A;648~A>G;790~G>A;850~T>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;3519~T>G;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;11861~A>G;12007~C>CTCG;12009~T>C;12391~C>T;14526~G>T;14777~C>G;15005~G>T;15165~C>G;15740~GTGT>G;15756~AT>A;16220~G>T;16402~C>G;**16490~A>T**;16648~A>G;16734~A>C;16818~T>G;16825~A>G;16831~G>T;16832~T>G;17194~C>T;17344~G>A;17488~G>GAGGAGAATGGTGCCCAAGA;18109~G>C;18202~G>A;18214~T>C;18259~T>G;18385~T>G;18399~G>T;18848~G>C;19008~A>C;19290~C>T;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19508~G>A;19518~C>G;19528~G>A;19589~G>T;19760~TA>T;19783~A>G;19897~C>G;19995~C>T;20226~G>A;20288~CCTTTCTTT>C;20389~TTC>*;20435~T>C;20717~T>C;20728~G>T;20753~G>A;20818~T>C;21618~C>T;21757~C>T	3
HG02572 (Child)	Hap 1 (*1C+**9809~G>A**): 364~C>A;648~A>G;734~ATT>A;790~G>A;850~T>A;851~A>T;3519~T>G;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10839~A>G;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12506~AT>A;15496~C>T;15756~A>AT;18728~C>T	
	Hap 2 (*1C): 790~G>A;3642~T>G;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20373~TTCTTTCTTTCTTTCTTTC>T	

Unphased variants: 3468~T>A;4227~T>C		
HG03577 (Father)	Hap 1 (*1C+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;1317~C>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12506~AT>A;15496~C>T;18728~C>T;20288~CCTTT>C	
	Hap 2 (*1C+**9809~G>A**+**16490~A>T**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12502~A>AT;12506~A>T;14260~T>TGGG;14526~G>T;14941~C>T;15005~G>T;15165~C>G;15740~GTGT>G;15756~AT>A;16220~G>T;16402~C>G;**16490~A>T**;16648~A>G;16734~A>C;17194~C>T;17344~G>A;18109~G>C;18202~G>A;18214~T>C;18259~T>G;18385~T>G;18399~G>T;18848~G>C;19008~A>C;19290~C>T;19302~C>G;19362~A>C;19374~TA>T;19383~T>C;19386~G>A;19508~G>A;19518~C>G;19528~G>A;19534~C>T;19545~C>A;19589~G>T;19760~TAA>T;19783~A>G;19995~C>T;20288~CCTTT>C;20370~T>C;20385~TTCTTTC>T;20717~T>C;20728~G>T;20753~G>A;20818~T>C;21491~C>T;21618~C>T;21757~C>T	2
HG03578 (Mother)	Hap 1 (*1C+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;15496~C>T;15756~A>AT	
	Hap 2 (*1C+**4682~A>G**): 286~T>C;364~C>A;648~A>G;790~G>A;850~T>A;1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;3519~T>G;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;6778~G>A;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;11861~A>G;12007~C>CTCG;12009~T>C;12391~C>T;15496~C>T;15756~A>AT;18728~C>T;19897~C>G;20317~C>T;20318~TTTTC>T;20318~T>C;20370~T>C;20377~TTCTTTCTTTCTTTC>T;20381~TTCTTTCTTTC>*	
HG03579 (Child)	Hap 1 (*1C+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;1317~C>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12506~AT>A;15496~C>T;18728~C>T;20288~CCTTT>C	
	Hap 2 (*1C+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;15496~C>T;15756~A>AT	
HG03514 (Mother)	Hap 1 (*1C+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;2961~C>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12321~T>C;12506~AT>A;14260~T>TGGG;15496~C>T;20381~T>TTC;20389~TTC>T	50

	Hap 2 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;14260~T>TGGG;14614~G>A;15496~C>T;20314~TTTCTTTTCTTTCTTTC>T;20381~T>TTC;20389~TTC>T	
HG03515 (Father)	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;14260~T>TGGG;14614~G>A;15496~C>T;20314~TTTCTTTTCTTTCTTTC>T;20381~T>TTC;20389~TTC>T	
	Hap 2 (**1C*): 133522479~T>G;133523152~G>A;133525299~C>T;133526023~G>A;133527362~T>TC;133528756~C>T;133529140~G>A;133531025~T>C;133533893~C>T;133534029~A>G;133534867~TA>T;133537858~C>T;133542735~TTCTTTCTTTCTTTCTTTC>T	
HG03516 (Child)	Hap1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;2961~C>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12321~T>C;12506~AT>A;14260~T>TGGG;15496~C>T;20381~T>TTC;20389~TTC>T	
	Hap2 (**1C*): 133522479~T>G;133523152~G>A;133525299~C>T;133526023~G>A;133527362~T>TC;133528756~C>T;133529140~G>A;133531025~T>C;133533893~C>T;133534029~A>G;133534867~TA>T;133537858~C>T;133542735~TTCTTTCTTTCTTTCTTTC>T	
NA19222 (Mother)	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;2803~C>T;5000~T>TC;7369~G>A;8140~T>C;8515~G>A;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;12009~T>TGTC;12506~AT>A;12891~C>G;15496~C>T;20288~CCTTTCTTT>C;20381~T>TTC;20389~TTC>T	
	Hap 2 (**1C*+**9809~G>A**+**15271~G>C**): 364~C>A;648~A>G;790~G>A;850~T>A;2794~C>G;3519~T>G;3690~C>G;4227~T>C;5000~T>TC;5582~G>A;6778~G>A;8663~T>C;8713~A>G;8957~T>A;9527~G>T;9570~C>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>C;12506~A>AT;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16648~A>G;16822~TTTATTTTTG>T;16989~T>C;17213~A>C;17633~T>G;17752~C>T;18259~T>G;18385~T>G;18665~C>A;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20226~G>A;20381~T>TTC;20435~T>C;20717~T>C;20928~G>T;21491~C>T;21618~C>T;21757~C>T	23

NA19223 (Father)	Hap 1 (**1C*): 8663~T>C;11531~C>T;11667~A>G;15496~C>T;20318~T>TTTTC	
	Hap 2 (**1C*): 11531~C>T;12502~A>AT;12506~A>T;14260~T>TGGG;15496~C>T;20318~T>TTTTC;20387~C>T;20389~T>C;20391~C>CTCTTTCTT;20391~C>T	
NA19221 (Child)	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;2803~C>T;5000~T>TC;7369~G>A;8140~T>C;8515~G>A;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;12009~T>TGTC;12506~AT>A;12891~C>G;15496~C>T;20288~CCTTTCTTT>C;20381~T>TTC;20389~TTC>T	
	Hap 2 (**1C*): 8663~T>C;11531~C>T;11667~A>G;15496~C>T;20318~T>TTTTC	
NA18909 (Mother)	Hap 1 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;20389~T>C;20391~C>T	
	Hap 2 (**1C*): 10260~T>C;11531~C>T;15496~C>T;20314~TTTCTTTTCTTTC>T;20381~T>TTC;20389~TTC>T;20391~C>CTCTTTCTT	
NA18910 (Father)	Hap 1 (**1C*+**3739~G>C**+**9809~G>A**+**15271~G>C**): 364~C>A;648~A>G;790~G>A;850~T>A;1855~A>C;3468~T>A;3519~T>G;**3739~G>C**;4007~T>C;4227~T>C;5000~T>TC;6778~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;9886~T>C;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16648~A>G;16654~T>C;16989~T>C;17213~A>C;17633~T>G;18259~T>G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20226~G>A;20366~CT>C;20389~T>C;20391~C>T;20435~T>C;20717~T>C;20928~G>T;21345~A>C;21491~C>T;21618~C>T;21757~C>T	24
	Hap 2 (**1C*+**3739~G>C**+**9809~G>A**+**15271~G>C**): 364~C>A;648~A>G;790~G>A;850~T>A;1855~A>C;3468~T>A;3519~T>G;**3739~G>C**;4007~T>C;4227~T>C;5000~T>TC;6778~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14466~G>T;14499~C>T;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16648~A>G;16654~T>C;16824~TATTTTTG>T;16989~T>C;17213~A>C;17633~T>G;18259~T>G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20297~C>CT;20317~CT>C;20366~CT>C;20389~T>C;20391~C>T;20435~T>C;20717~T>C;20928~G>T;21345~A>C;21491~C>T;21618~C>T;21757~C>T	

NA18911 (Child)	Hap 1 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;20389~T>C;20391~C>T	
	Hap 2 (**1C*+**3739~G>C**+**9809~G>A**+**15271~G>C**): 364~C>A;648~A>G;790~G>A;850~T>A;1855~A>C;3468~T>A;3519~T>G;**3739~G>C**;4007~T>C;4227~T>C;5000~T>TC;6778~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;12891~C>G;13360~C>A;13501~G>A;13671~G>A;13935~T>C;14048~C>T;14213~A>G;14466~G>T;14499~C>T;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;16147~C>T;16210~A>G;16234~T>C;16287~A>G;16302~T>C;16648~A>G;16654~T>C;16824~TATTTTTG>T;16989~T>C;17213~A>C;17633~T>G;18259~T>G;18385~T>G;18749~TAAGAA>T;18848~G>C;19008~A>C;19141~C>T;19415~C>CTCTTAAT;19416~A>G;19528~G>A;19534~C>T;19545~C>A;19760~TA>T;20297~C>CT;20317~CT>C;20366~CT>C;20389~T>C;20391~C>T;20435~T>C;20717~T>C;20928~G>T;21345~A>C;21491~C>T;21618~C>T;21757~C>T	
HG03168 (Mother)	Hap 1 (**1C*+**4963~G>T**): 790~G>A;851~A>T;2530~T>G;**4963~G>T**;5000~T>TC;6069~C>T;6394~C>T;6778~G>A;8663~T>C;8713~A>G;9945~T>C;10839~A>G;11667~A>G;12009~T>TGTC;12506~AT>A;14259~GT>G;15496~C>T;20389~TTC>T;20444~GTTTC>G	8
	Hap 2 (**1C*): 364~C>A;648~A>G;790~G>A;1846~C>T;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;8663~T>C;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;12891~C>G;14260~T>TGGG;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;17432~G>A;18728~C>T;20288~CCTTTCTTT>C;20366~CT>C;20376~TTTC>T;20389~TTC>T	
HG03169 (Father)	Hap 1 (**1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;14259~GT>G;15496~C>T;15756~A>AT;18728~C>T;20389~TTC>T	
	Hap 2 (**1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12007~C>CTCG;12009~T>C;12506~AT>A;14260~T>TGGG;15120~C>T;15496~C>T;16624~C>T;16855~T>C;17112~C>T;20288~CCTTTCTTT>C;20314~TTTCTTTTCTTTC>T;20366~CTTTTCTTTCTTTCTTTCT>C;20385~T>C;20387~C>T;20389~T>C;20391~C>T	
HG03170 (Child)	Hap 1 (**1C*+**4963~G>T**): 790~G>A;851~A>T;2530~T>G;**4963~G>T**;5000~T>TC;6069~C>T;6394~C>T;6778~G>A;8663~T>C;8713~A>G;9945~T>C;10839~A>G;11667~A>G;12009~T>TGTC;12506~AT>A;14259~GT>G;15496~C>T;20389~TTC>T;20444~GTTTC>G	

	Hap 2 (*1C*+**9809~G>A**): 734~ATT>A;790~G>A;851~A>T;5000~T>TC;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10132~C>T;10151~C>T;10152~A>G;10830~C>T;11667~A>G;12009~T>TGTC;12506~AT>A;14259~GT>G;15496~C>T;15756~A>AT;18728~C>T;20389~TTC>T	
HG02702 (Father)	Hap 1 (*1C*): 117~T>G;364~C>A;648~A>G;790~G>A;850~T>A;2937~C>T;3519~T>G;3661~G>A;5000~T>TC;6394~C>T;6778~G>A;8663~T>C;11531~C>T;11667~A>G;15496~C>T;15756~A>AT;18728~C>T;20385~T>*;20391~C>CTT	
	Hap 2 (*1C*): 790~G>A;3468~T>A;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;11667~A>G;12009~T>TGTC;12505~TA>T;12506~AT>A;15120~C>T;15496~C>T;16624~C>T;16855~T>C;17112~C>T;20288~CCTTTCTTT>C	
HG02703 (Mother)	Hap 1 (*1C*+**4682~A>G**): 790~G>A;1363~C>T;1364~A>G;1369~T>C;2136~A>G;2217~T>C;2739~C>T;3378~G>C;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;6778~G>A;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;11861~A>G;12007~C>CTCG;12009~T>C;12391~C>T;15496~C>T;19897~C>G;20356~TTC>T	42
	Hap 2 (*1C*): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;3642~T>G;4227~T>C;5000~T>TC;6916~C>G;8317~C>T;8663~T>C;8883~A>C;11667~A>G;12009~T>TGTC;12506~AT>A;15120~C>T;15496~C>T;15756~A>AT;16624~C>T;16855~T>C;17112~C>T;18728~C>T;20288~CCTTTCTTT>C;20314~TTTCTTTTC>T;20373~TTCTTTCTTTCTTTCTTTC>T;20381~TTCTTTCTTTC>*	
HG02704 (Child)	Hap 1 (*1C*): 117~T>G;364~C>A;648~A>G;790~G>A;850~T>A;2937~C>T;3519~T>G;3661~G>A;5000~T>TC;6394~C>T;6778~G>A;8663~T>C;11531~C>T;11667~A>G;15496~C>T;15756~A>AT;18728~C>T;20385~T>*;20391~C>CTT	
	Hap 2 (*1C*+**4682~A>G**): 790~G>A;1363~C>T;1364~A>G;1369~T>C;2136~A>G;2217~T>C;2739~C>T;3378~G>C;4104~A>G;**4682~A>G**;5000~T>TC;6394~C>T;6778~G>A;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;11861~A>G;12007~C>CTCG;12009~T>C;12391~C>T;15496~C>T;19897~C>G;20356~TTC>T	
HG02860 (Father)	Hap 1 (*1C*+**4682~A>G**+**9809~G>A**): 364~C>A;648~A>G;734~ATT>A;790~G>A;850~T>A;851~A>T;3519~T>G;**4682~A>G**;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8845~T>C;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10830~C>T;10839~A>G;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12391~C>T;12506~AT>A;15496~C>T;15756~A>AT;18728~C>T;20288~CCTTT>C;20381~TTCTTTCTTTC>T	

	Hap 2 (**1C*): 790~G>A;5000~T>TC;8663~T>C;11667~A>G;14260~T>TGGG;15496~C>T;15756~A>AT;17488~G>GAGGAGAATGGTGCCCAAGA;18728~C>T;20288~CCTTT>C;20385~TTCTTTC>*	
HG02861 (Mother)	Hap 1 (**1C*): 790~G>A;2136~A>G;5000~T>TC;5828~C>T;8663~T>C;11667~A>G;15496~C>T;15756~A>AT;18728~C>T;20381~TTCTTTCTTTC>T	
	Hap 2 (**1C*+**3739~G>C**+**4682~A>G**): 364~C>A;648~A>G;790~G>A;850~T>A;3468~T>A;3519~T>G;**3739~G>C**;4227~T>C;**4682~A>G**;5000~T>TC;6394~C>T;8663~T>C;8845~T>C;10830~C>T;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>C;12391~C>T;12678~T>A;12713~A>G;12807~C>A;12891~C>G;12971~T>TA;13091~C>A;13360~C>A;13501~G>A;13671~G>A;13701~G>A;13935~T>C;14048~C>T;14260~T>TGGG;14363~ATGGTTGGATGGATGGGTGGGTGGGTGGATGGATGGG>A;15005~G>T;**15271~G>C**;15308~T>A;15740~GTGT>G;15756~AT>A;19337~C>T;20379~CTT>C;20389~TTC>*;20448~C>G;21067~G>A	2
HG02862 (Child)	Hap 1 (**1C*+**4682~A>G**+**9809~G>A**): 364~C>A;648~A>G;734~ATT>A;790~G>A;850~T>A;851~A>T;3519~T>G;**4682~A>G**;5000~T>TC;6394~C>T;8140~T>C;8515~G>A;8663~T>C;8825~C>T;8845~T>C;8957~T>A;9527~G>T;9737~G>A;**9809~G>A**;10830~C>T;10839~A>G;11061~C>A;11667~A>G;12007~C>CTCG;12009~T>TGTC;12009~T>C;12391~C>T;12506~AT>A;15496~C>T;15756~A>AT;18728~C>T;20288~CCTTT>C;20381~TTCTTTCTTTC>T	
	Hap 2 (**1C*): 790~G>A;2136~A>G;5000~T>TC;5828~C>T;8663~T>C;11667~A>G;15496~C>T;15756~A>AT;18728~C>T;20381~TTCTTTCTTTC>T	
Unphased variants: 1363~C>T;1364~A>G;1369~T>C;2217~T>C;2739~C>T;3378~G>C;4104~A>G;6778~G>A;11861~A>G;19897~C>G		

Variants bolded in blue indicate additional variants defining possible novel *CYP2E1* star alleles/sub-alleles. Variants bolded in black indicate variants defining known *CYP2E1* star alleles

Hap - Haplotype

Supplementary Table 4: The distribution of *NAT2* star alleles with unknown acetylation effects in SSA compared with various global populations

Star alleles	Defining variants	Consequences	Allele frequencies (%)*					
			SSA **(n=792)**	African American/Afro-Caribbean **(n=157)**	AMR **(n=346)**	EAS **(n=504)**	EUR **(n=503)**	SAS **(n=488)**
**5L*	rs45477599 rs1801280 rs1799929 rs1208	Missense (L24I) Missense (I114T) Synonymous (L161L) Missense (K268R)	0.4	0	0	0	0	0
**5N*	rs1801280 rs139351995 rs1799929 rs1208	Missense (I114T) Missense (I158L) Synonymous (L161L) Missense (K268R)	0.4	0.3	0	0	0	0
**5P*	rs1041983 rs1801280 rs1799929 rs79050330 rs1799930 rs1208	Synonymous (Y94Y) Missense (I114T) Synonymous (L161L) Missense (T193M) Missense (R197Q) Missense (K268R)	0	0	0	0.2	0	0
**5W*	rs1801280 rs146405047 rs1799929 rs1208	Missense (I114T) Synonymous (N118N) Synonymous (L161L) Missense (K268R)	0.1	0.3	0	0	0	0
**6H*	rs1041983 rs1799930 rs55700793	Synonymous (Y94Y) Missense (R197Q) Stop gained (K256E)	1.1	0.6	0.1	0	0	0
**6K*	rs1041983 rs1799930 rs138707146	Synonymous (Y94Y) Missense (R197Q) Missense (P213L)	0.2	0	0	0	0	0
**6L*	rs1041983 rs45532639 rs1799930	Synonymous (Y94Y) Synonymous (D115D) Missense (R197Q)	0.1	0	0	0	0	0.1

**6O*	rs1041983 rs1799930 rs56393504	Synonymous (Y94Y) Missense (R197Q) Missense (V280M)	2.7	4.8	0	0.1	0	0
**6P*	rs12720065 rs1799930	Missense (L135V) Missense (R197Q)	1.5	0.6	0	0	0	0
**7D*	rs1801279 rs1799930 rs1799931	Missense (R64Q) Missense (R197Q) Missense (G286E)	0.5	1.3	0	0	0	0
**12E*	rs1041983 rs79050330 rs1208	Synonymous (Y94Y) Missense (T193M) Missense (K268R)	0.3	0.3	0	0	0	0
**12F*	rs56387565 rs1208	Missense (Y208H) Missense (K268R)	0.1	0	0	0	0	0
**12G*	rs45618543 rs1208	Missense (E203D) Missense (K268R)	0.9	0.6	0.1	0	0	0
**12H*	rs12720065 rs1208	Missense (L135V) Missense (K268R)	1.5	0.6	0.1	0	0	0
**12N*	rs149283608 rs1208	Missense (N41Y) Missense (K268R)	0.1	0	0	0	0	0
**12O*	rs72466456 rs45618543 rs1208	Missense (I10T) Missense (E203D) Missense (K268R)	0	0.6	0	0	0	0
13F	rs1041983 rs56393504	Synonymous (Y94Y) Missense (V280M)	0.1	0	0	0	0	0
**14H*	rs1801279 rs1041983 rs45518335	Missense (R64Q) Synonymous (Y94Y) Missense (P228L)	0.4	0.6	0	0	0	0
**14L*	rs200893121 rs1801279 rs1041983	Missense (I3V) Missense (R64Q) Synonymous (Y94Y)	0.1	0	0	0	0	0
**24*	rs12720065	Missense (L135V)	0.1	0	0	0	0	0
**26*	rs868725509	Missense (I270T)	0.3	0	0	0	0	0
**27*	rs375746304	Stop gain (R197*)	0.2	0	0	0	0	0
**0*		Null deletion	0	0	0	0	0	0.2

**4x3*		Duplication	0	0	0	0	0.1	0
**4x2*		Duplication	0	0	0	0.1	0	0
**6Ax2*		Duplication	0	0	0	0.1	0	0

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants

*Allele frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

Supplementary Table 5: The distribution of *NAT2* star alleles with unknown acetylation effects across SSA populations in this study

Star allele	Allele frequencies (%)*								
	All **(n=792)**	BFA **(n=32)**	ESN **(n=98)**	GWD **(n=113)**	GHA **(n=26)**	LWK **(n=99)**	MSL **(n=85)**	SEB **(n=200)**	YRI **(n=108)**
**5L*	0.4	0	0	0	0	1.5	0	4	0
**5N*	0.4	0	0.5	0	0	0	0.6	4	0.5
**5W*	0	0	0	0.4	0	0	0	0	0
**6H*	0.1	1.6	0.5	0.4	0	0.5	0.6	0.5	0.5
**6K*	1.1	0	0	0.9	0	0	0.6	0	0
**6L*	0.2	0	0	0.4	0	0	0.6	0	0
**6O*	2.7	7.8	3.0	3.1	1.9	1.0	4.1	1.2	3.7
**6P*	1.5	0	0	0	0	3.0	0	4	0.5
**7D*	0.5	0	1.0	1.8	0	0	1.2	0	0
**12E*	0.3	0	0	0.9	0	0	0	0	0.9
**12F*	0.1	0	0	0	0	0.5	0	0	0
**12G*	0.9	1.6	1.5	0.9	1.9	0	1.2	2	1.4
**12H*	1.5	0	1.5	4.4	1.9	2.5	0	2	0.9
**12N*	0.1	0	0.5	0	0	0	0	0	0
**14H*	0.4	0	0	0	0	1.5	0	0	1.4
**14L*	0.1	0	0	0	0	0	0	0	0.5
**26*	0.3	0	0	0	0	0	0	5	0
**27*	0.2	0	0	0	0	0	0	3	0

BFA – Burkina Faso participants; ESN – Esan in Nigeria; GWD – Gambian in Western Division, Mandinka; GHA – Ghanaian participants; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; SEB – Southeastern Bantu in South Africa; YRI – Yoruba in Ibadan

*Allele frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

Supplementary Table 6: Examples of trio data used to infer core *NAT1* (NG_012245.2) and *NAT2* (NG_012246.1) variants defining novel haplotypes

Coriell ID	Haplotypes
NAT1	
HG02461 (Father)	Hap 1 (*3): **57674~A>T**
	Hap 2 (*4): **57674~A>T**;**57681~A>C**;57777~T>G;58302~C>T
HG02462 (Mother)	Hap 1 (*4+**57068~T>C**) : **57068~T>C**;**57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*3): **57674~A>T**;58302~C>T
HG02463 (Child)	Hap 1 (*3): **57674~A>T**
	Hap 2 (*4+**57068~T>C**): **57068~T>C**;**57674~A>T**;**57681~A>C**;57777~T>G
Unphased variants: 55658~G>A;55851~GT>G	
HG02973 (Father)	Hap 1 (*4): **57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*10+**18223111~A>AAAT**): **18223111~A>AAAT**
HG02974 (Mother)	Hap 1 (*4): **57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*4): **57674~A>T;57681~A>C**;57777~T>G
HG02975 (Child)	Hap 1 (*4): **57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*4): **57674~A>T;57681~A>C**;57777~T>G
NA18516 (Father)	Hap 1 (*4+**57277~G>T**): **57277~G>T**;**57674~A>T**; **57681~A>C**;57777~T>G
	Hap 2 (*4): **57674~A>T**;**57681~A>C**;57777~T>G
NA18517 (Mother)	Hap 1 (*4+**57650~A>AAAT**): 55658~G>A;55851~GT>G;**57650~A>AAAT**;**57674~A>T**;**57681~A>C**;57777~T>G;57841~T>C
	Hap 2 (*10):
NA18515 (Child)	Hap 1 (*4+**57277~G>T**): **57277~G>T**;**57674~A>T**; **57681~A>C**;57777~T>G
	Hap 2 (*4+**57650~A>AAAT**): 55658~G>A;55851~GT>G;**57650~A>AAAT**;**57674~A>T**;**57681~A>C**;57777~T>G;57841~T>C
HG03311 (Father)	Hap 1 (*3): 55539~G>C;55658~G>A;55851~GT>G;56551~A>T;**57674~A>T**;58302~C>T
	Hap 2 (*10+**57650~A>AAAT**+**57681~A>C**): **57650~A>AAAT**;**57681~A>C**;57777~T>G;58321~A>T
HG03309 (Mother)	Hap 1 (*3): 55658~G>A; **57674~A>T**;55851~GT>G;58302~C>T;58371~A>G
	Hap 2 (*4): **57674~A>T**;**57681~A>C**;57777~T>G
HG03310 (Child)	Hap 1 (*3): 55658~G>A; **57674~A>T**;55851~GT>G;58302~C>T;58371~A>G
	Hap 2 (*3): 55658~G>A;55851~GT>G;58302~C>T;58371~A>G
	Hap 1 (*4): **57674~A>T;57681~A>C**;57777~T>G

HG02943 (Mother)	Hap 2 (*4*+**56607~T>G**): 55658~G>A;55851~GT>G;56033~G>A;**56607~T>G**;**57674~A>T**;**57681~A>C**;57777~T>G;58272~G>A;58302~C>T
HG02944 (Father)	Hap 1 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
HG02945 (Child)	Hap 1 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
HG02946 (Mother)	Hap 1 (*10*+**56777~G>A**): 55658~G>A;55851~GT>G;56211~T>C;56309~T>A;**56777~G>A**;58302~C>T
	Hap 2 : 55658~G>A;55851~GT>G;58275~A>G
HG02947 (Father)	Hap 1 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
HG02948 (Child)	Hap 1 (*10*+**56777~G>A**): 55658~G>A;55851~GT>G;56211~T>C;56309~T>A;**56777~G>A**;58302~C>T
	Hap 2 (*4*): **57674~A>T**;**57681~A>C**;57777~T>G
HG02464 (Father)	Hap 1 (*10*): 55658~G>A;58275~A>G
	Hap 2 (*10*): 55658~G>A;55851~GT>G;56551~A>T;58228~A>C;58233~T>C;58321~A>T
HG02465 (Mother)	Hap 1 (*3*+**57650~A>AAATAATAAT**): 55658~G>A;55851~GT>G;**57650~A>AAATAATAAT**;**57674~A>T**;57777~T>G
	Hap 2 (*10*):
HG02466 (Child)	Hap 1 (*10*): 55658~G>A;58275~A>G
	Hap 2 (*3*+**57650~A>AAATAATAAT**): 55658~G>A;55851~GT>G;**57650~A>AAATAATAAT**;**57674~A>T**;57777~T>G
Unphased variant: 58302~C>T	
HG02620 (Father)	Hap 1 (*10*): 58302~C>T;58371~A>G
	Hap 2 (*10*+**57025~C>T**): **57025~C>T**
HG02621 (Mother)	Hap 1 (*15*): **57145~C>T**;**57674~A>T**;**57681~A>C**;57777~T>G
	Hap 2 (*10):
HG02622 (Child)	Hap 1 (*10*): 55658~G>A;58275~A>G
	Hap 2 (*10*+**57025~C>T**): **57025~C>T**
Unphased variants: 55658~G>A;55851~GT>G	
NAT2	
HG03270 (Mother)	Hap 1 (*6A*+**14007~G>C**): 484~G>C;707~T>A;797~G>A;908~G>T;930~G>A;999~C>T;1240~G>C;1695~A>G;1904~T>C;2260~C>CAAA;2299~A>G;2705~G>T;3476

	~TTCTGCTAATC>T;4884~A>G;6755~G>C;7386~A>G;7827~A>G;8246~G>A;10821~T>C;11585~C>G;12122~C>T;12905~T>C;**14007~G>C**;**14041~C>T**;**14349~G>A**;**14562~G>A**;15154~G>A;15551~A>G;15612~C>T
	Hap 2 (**6A*): 484~G>C;1099~G>T;1240~G>C;1904~T>C;4884~A>G;7386~A>G;7827~A>G;8246~G>A;9707~C>CT;10821~T>C;11585~C>G;12122~C>T;12905~T>C;**14041~C>T**;**14349~G>A**;**14562~G>A**;15154~G>A;15551~A>G;15612~C>T
HG03271 (Father)	Hap 1 (**6A*): -203~C>T;484~G>C;908~G>T;930~G>A;999~C>T;1240~G>C;1695~A>G;1904~T>C;2185~G>A;2260~C>CAAA; 4884~A>G;7386~A>G;7827~A>G;8246~G>A;9707~C>CTT;10821~T>C;11585~C>G;12122~C>T;12905~T>C;**14041~C>T**;**14349~G>A**;**14562~G>A**;15154~G>A;15551~A>G;15612~C>T
	Hap 2 (**5B*): 447~C>T;484~G>C;1240~G>C;1904~T>C;2260~C>CAAA;2299~A>G;2910~A>G;3952~C>CAG;4480~C>CAATT;4884~A>G;5898~T>C;6838~G>T;7168~T>C;7386~A>G;7756~G>A;7804~T>G;7827~A>G;8055~G>A;8246~G>A;8315~G>C;8392~CCTT>C;9001~T>C;9655~C>G;10613~G>A;10618~T>C;11185~T>C;11585~C>G;12363~C>T;12687~T>C;12905~T>C;13269~A>G;13526~C>T;**14100~T>C**;**14240~C>T**;15612~C>T
HG03272 (Child)	Hap 1 (**6A*+**14007~G>C**): 484~G>C;707~T>A;797~G>A;908~G>T;930~G>A;999~C>T;1240~G>C;1695~A>G;1904~T>C;2260~C>CAAA;2299~A>G;2705~G>T;3476~TTCTGCTAATC>T;4884~A>G;6755~G>C;7386~A>G;7827~A>G;8246~G>A;10821~T>C;11585~C>G;12122~C>T;12905~T>C;**14007~G>C**;**14041~C>T**;**14349~G>A**;**14562~G>A**;15154~G>A;15551~A>G;15612~C>T
	Hap 2 (**6A*): -203~C>T;484~G>C;908~G>T;930~G>A;999~C>T;1240~G>C;1695~A>G;1904~T>C;2185~G>A;2260~C>CAAA; 4884~A>G;7386~A>G;7827~A>G;8246~G>A;9707~C>CTT;10821~T>C;11585~C>G;12122~C>T;12905~T>C;**14041~C>T**;**14349~G>A**;**14562~G>A**;15154~G>A;15551~A>G;15612~C>T
Unphased variants: 2379~A>C;2942~T>C;3695~A>G;3964~T>A;6563~A>T;8172~A>G	

Variants bolded in blue indicate additional variants defining possible novel *NAT1* and *NAT2* star alleles/sub-alleles. Variants bolded in black indicate variants defining known *NAT1* and *NAT2* star alleles

Hap – Haplotype

Supplementary Table 7: The distribution of *NAT1* diplotypes and NAT1 predicted phenotypes in SSA compared with various global populations

Haplotype	Acetylation effect	Diplotype frequencies (%)*					
		SSA **(n=762)**	African American/ Afro-Caribbean **(n=152)**	AMR **(n=344)**	EAS **(n=461)**	EUR **(n=477)**	SAS **(n=450)**
**10/*10*	Rapid	27.9	15.1	15.1	25.2	5.5	11.6
**11A/*3*	Intermediate	0	0	0	0	0	0.4
**11A/*4*	Intermediate	0	0.7	2.0	0.7	3.6	5.3
**3/*10*	Intermediate	2.5	0.7	0.3	1.5	0.8	3.3
**4/*10*	Intermediate	47.1	49.3	38.1	45.3	29.8	40.2
**10/*15*	Intermediate	0.1	0	0.3	0	0.2	0
**10/*27*	Intermediate	1.3	3.3	0.3	0	0	0
**10/*17*	Intermediate	0	0.7	0	0	0.2	0
**10/14A*	Intermediate	0	0.7	0.6	0	1.0	0
**10/*20*	Intermediate	0	0.7	0	0	0	0
**10/*22*	Intermediate	0	0	0.3	0	0	0
**3/*4*	Slow	2.1	2.6	2.6	2.2	3.4	3.6
**4/*4*	Slow	17.6	22.4	39.8	21.0	52.0	34.4
**27/*4*	Slow	0.9	2.6	0	0	0	0
**15/*4*	Slow	0.1	0.7	0.6	0	1.5	0
**3/*3*	Slow	0	0.7	0	0	0	0.7
**10x2/*4*	Unknown	0	0	0	0	0	0.2
**10x2/*3x2*	Unknown	0	0	0	0	0	0.2
**17/*4*	Unknown	0	0	0	0	0.8	0
**22/*4*	Unknown	0	0	0	0	0.8	0
**15/*22*	Unknown	0	0	0	0	0.2	0
**10/*26B*	Unknown	0.3	0	0	0	0	0

**26B/*4*	Unknown	0.1	0	0	0	0	0
**16/4*	Unknown	0	0	0	3.7	0	0
**10x2/*4x2*	Unknown	0	0	0	0.2	0	0
**16/*3*	Unknown	0	0	0	0.2	0	0
**0/*0*	Unknown	0	0	0	0	0.2	0

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants

*Diplotype frequencies calculated with excluding individuals with novel haplotypes and/or unresolved diplotypes

Supplementary Table 8: The distribution of *NAT2* diplotypes and NAT2 predicted phenotypes in SSA compared with various global populations

Diplotypes	Acetylation effect	Diplotype frequencies (%)					
		SSA **(n=791)**	African American/Afro-Caribbean **(n=157)**	AMR **(n=345)**	EAS **(n=504)**	EUR **(n=503)**	SAS **(n=487)**
**4/*4*	Rapid	1.0	1.9	12.5	30.0	6.6	4.9
**4/*12A*	Rapid	1.4	0.6	0.9	0.2	0.2	1.0
**4/*12B*	Rapid	0.1	0.0	0.0	0.0	0.0	0.0
**4/*13A*	Rapid	2.8	2.5	0.6	0.8	0.0	0.0
**12A/*12A*	Rapid	1.5	0.0	0.0	0.0	0.0	0.0
**12A/*12B*	Rapid	0.5	0.0	0.0	0.0	0.0	0.0
**12A/*13A*	Rapid	1.1	0.6	0.0	0.0	0.0	0.0
**13A/*13A*	Rapid	1.9	2.5	0.0	0.0	0.0	0.0
**4/*5A*	Intermediate	0.1	0.6	0.0	0.0	0.6	0.2
**4/*5B*	Intermediate	5.1	5.1	21.7	3.0	19.5	11.7
**4/*5C*	Intermediate	1.3	1.9	0.3	0.2	0.4	1.2
**4/*6A*	Intermediate	2.0	7.0	8.7	21.8	12.5	14.4
**4/*6B*	Intermediate	0.0	0.0	0.0	0.0	0.4	0.0
**4/*7B*	Intermediate	0.3	1.3	2.0	5.0	0.8	0.8
**4/*10*	Intermediate	0.0	0.0	0.0	0.2	0.0	0.0
**4/*14A*	Intermediate	0.8	0.6	0.0	0.0	0.0	0.0
**4/*14B*	Intermediate	2.1	0.6	0.3	0.0	0.0	0.0
**5A/*13A*	Intermediate	0.1	0.0	0.0	0.0	0.0	0.0
**5B/*12A*	Intermediate	4.8	2.5	1.2	0.0	0.2	2.1
**5B/*12B*	Intermediate	0.4	0.0	0.0	0.0	0.0	0.0
**5B/*13A*	Intermediate	3.8	3.2	0.3	0.0	0.0	0.0
**5C/*12A*	Intermediate	1.6	0.0	0.0	0.0	0.0	0.0
**5C/*12B*	Intermediate	0.4	0.0	0.0	0.0	0.0	0.0
**5C/*13A*	Intermediate	0.6	1.9	0.0	0.0	0.0	0.0

**5D/*13A*	Intermediate	0.4	0.0	0.0	0.0	0.0	0.0
**6A/*12A*	Intermediate	4.4	1.9	0.0	0.2	0.2	0.4
**6A/*12B*	Intermediate	0.4	0.0	0.0	0.0	0.0	0.0
**6A/*12C*	Intermediate	0.0	0.0	0.0	0.0	0.0	0.2
**6A/*13A*	Intermediate	4.3	4.5	0.6	0.2	0.0	0.0
**7B/*12A*	Intermediate	0.1	0.0	0.3	0.0	0.0	0.0
**7B/*12B*	Intermediate	0.1	0.6	0.6	0.0	0.2	0.0
**7B/*13A*	Intermediate	0.3	0.0	0.0	0.0	0.0	0.0
**12A/*14B*	Intermediate	0.8	1.3	0.0	0.0	0.0	0.0
**12B/*14B*	Intermediate	0.0	0.6	0.0	0.0	0.0	0.0
**13A/*14A*	Intermediate	0.1	0.0	0.0	0.0	0.0	0.0
**13A/*14B*	Intermediate	1.9	0.6	0.0	0.0	0.0	0.0
**5A/*5B*	Slow acetylator	0.3	0.0	0.0	0.0	1.4	0.2
**5A/*5C*	Slow acetylator	0.0	0.0	0.3	0.0	0.0	0.0
**5A/*6A*	Slow acetylator	0.0	0.0	0.6	0.0	1.2	0.2
**5A/*7B*	Slow acetylator	0.0	0.0	0.3	0.0	0.2	0.0
**5A/*14B*	Slow acetylator	0.1	0.0	14.2	0.0	0.0	0.0
**5B/*5B*	Slow acetylator	4.9	7.6	0.0	0.2	18.1	10.9
**5B/*5C*	Slow acetylator	2.8	1.9	0.6	0.0	2.0	1.6
**5B/*5E*	Slow acetylator	0.1	0.0	11.3	0.0	0.0	0.0
**5B/*6A*	Slow acetylator	7.5	11.5	4.3	3.0	21.5	20.9
**5B/*6C*	Slow acetylator	0.0	0.0	0.0	0.0	0.2	0.2
**5B/*7B*	Slow acetylator	1.0	0.0	0.0	1.0	2.2	4.1
**5B/*14A*	Slow acetylator	0.5	1.9	0.3	0.0	0.0	0.0
**5B/*14B*	Slow acetylator	3.2	0.6	0.0	0.0	0.2	0.0
**5B/*14G*	Slow acetylator	0.0	0.6	0.0	0.0	0.0	0.0
**5C/*5C*	Slow acetylator	0.8	0.0	0.0	0.0	0.2	0.0
**5C/*6A*	Slow acetylator	1.9	0.6	0.9	0.0	0.4	2.1
**5C/*6B*	Slow acetylator	0.0	0.0	0.3	0.0	0.0	0.0

**5C/*7B*	Slow acetylator	0.1	1.3	3.8	0.0	0.0	0.2
**5C/*14A*	Slow acetylator	0.3	0.0	0.0	0.0	0.0	0.0
**5C/*14B*	Slow acetylator	0.4	0.0	0.9	0.0	0.0	0.0
**5D/*5D*	Slow acetylator	0.1	0.0	4.1	0.0	0.0	0.0
**6A/*6A*	Slow acetylator	3.0	4.5	0.0	8.5	9.3	13.1
**6A/*6C*	Slow acetylator	0.0	0.0	0.0	0.0	0.0	0.2
**6A/*7B*	Slow acetylator	1.0	2.5	0.0	9.5	1.0	6.2
**6A/*14A*	Slow acetylator	0.9	0.0	0.0	0.0	0.0	0.0
**6A/*14B*	Slow acetylator	1.8	2.5	7.5	0.0	0.0	0.0
**6C/*7B*	Slow acetylator	0.0	0.0	0.0	0.0	0.0	0.2
**6P/*12A*	Slow acetylator	0.5	0.0	0.0	0.0	0.0	0.0
**7B/*7B*	Slow acetylator	0.1	1.3	0.0	16.1	0.2	2.1
**7B/*14A*	Slow acetylator	0.4	0.0	0.0	0.0	0.0	0.0
**14A/*14A*	Slow acetylator	1.0	0.0	0.0	0.0	0.0	0.0
**14A/*14B*	Slow acetylator	0.3	0.6	0.0	0.0	0.0	0.0
**14D/*14D*	Slow acetylator	0.4	0.0	0.0	0.0	0.0	0.0
**4/*5N*	Unknown	0.3	0.6	0.3	0.0	0.0	0.0
**4/*5P*	Unknown	0.0	0.0	0.0	0.0	0.0	0.4
**4/*6H*	Unknown	0.3	0.6	0.0	0.0	0.0	0.0
**4/*6O*	Unknown	0.1	0.6	0.0	0.0	0.0	0.2
**4/*6P*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**4/*12G*	Unknown	0.1	0.0	0.3	0.0	0.0	0.0
**4/*12H*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**4/*14H*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**4/*26*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**4/*27*	Unknown	0.4	0.0	0.0	0.0	0.0	0.0
**5A/*12H*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5A/*26*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5B/*5L*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0

**5B/*5W*	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
**5B/*6H*	Unknown	0.3	0.6	0.0	0.0	0.0	0.0
**5B/*6K*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5B/*6L*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5B/*6O*	Unknown	0.8	3.2	0.0	0.0	0.0	0.0
**5B/*6P*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**5B/*12E*	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
**5B/*12F*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5B/*12G*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5B/*12H*	Unknown	1.1	0.6	0.0	0.0	0.0	0.0
**5B/*14H*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5C/*5N*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5C/*5W*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5C/*6O*	Unknown	0.1	0.6	0.0	0.0	0.0	0.0
**5C/*6P*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**5C/*12H*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5C/*12G*	Unknown	0.4	0.0	0.0	0.0	0.0	0.0
**5C/*12N*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5L/*12A*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**5L/*13A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5L/*14A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5L/*14B*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**5N/*6A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5N/*12A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5N/*13A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**5N/*14B*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**6A/*6H*	Unknown	0.8	0.0	0.0	0.0	0.0	0.0
**6A/*6K*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**6A/*6L*	Unknown	0.0	0.0	0.0	0.0	0.2	0.0

*6A/*6O	Unknown	1.1	0.6	0.0	0.0	0.0	0.0
*6A/*6P	Unknown	0.5	0.6	0.0	0.0	0.0	0.0
*6A/*12E	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6A/*12G	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
*6A/*12H	Unknown	0.5	0.0	0.0	0.0	0.0	0.0
*6A/*14L	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6H/*12A	Unknown	0.4	0.0	0.0	0.0	0.0	0.0
*6H/*12B	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6H/*13A	Unknown	0.4	0.0	0.0	0.0	0.0	0.0
*6K/*14A	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6L/*14B	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6O/*6O	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
*6O/*7B	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6O/*12A	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
*6O/*12B	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
*6O/*12G	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6O/*12H	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6O/*13A	Unknown	1.6	1.9	0.0	0.0	0.0	0.0
*6O/*14A	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*6O/*14B	Unknown	0.5	0.6	0.0	0.0	0.0	0.0
*6P/*6P	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
*6P/*12B	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
*6P/*13A	Unknown	0.1	0.6	0.0	0.0	0.0	0.0
*6P/*14B	Unknown	0.4	0.0	0.0	0.0	0.0	0.0
*6P/*14H	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
*7B/*12G	Unknown	0.1	0.0	0.3	0.0	0.0	0.0
*7D/*7D	Unknown	0.5	1.3	0.0	0.0	0.0	0.0
*12A/*12G	Unknown	0.1	0.6	0.0	0.0	0.0	0.0
*12A/*12H	Unknown	0.1	0.6	0.0	0.0	0.0	0.0

**12A/*14H*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**12A/*26*	Unknown	0.4	0.0	0.0	0.0	0.0	0.0
**12B/*12G*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**12C/*12E*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**12E/*13A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**12E/*14A*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**12G/*13A*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**12G/*14A*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**12G/*14B*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**12H/*13A*	Unknown	0.5	0.0	0.0	0.0	0.0	0.0
**12O/*12O*	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
**13A/*14E*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**13A/*14H*	Unknown	0.3	0.6	0.0	0.0	0.0	0.0
**13F/*14D*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**14B/*14B*	Unknown	0.3	0.0	0.0	0.0	0.0	0.0
**14B/*14H*	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
**14B/*6O*	Unknown	0.0	0.6	0.0	0.0	0.0	0.0
**24/*24*	Unknown	0.1	0.0	0.0	0.0	0.0	0.0
**4x2/*6Ax2*	Unknown	0.0	0.0	0.0	0.0	0.0	0.2
**4/*4x3*	Unknown	0.0	0.0	0.0	0.2	0.0	0.0
**0/*0*	Unknown	0.0	0.0	0.0	0.0	0.2	0.0

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants

*Diplotype frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

Supplementary Table 9: The distribution of *GSTM1* diplotypes and the associated GSTM1 enzyme activity phenotypes in SSA compared with various global populations

Diplotype	Enzyme activity	Diplotype frequencies (%)					
		SSA **(n=765)**	African American/ Afro-Caribbean **(n=157)**	AMR **(n=340)**	EAS **(n=503)**	EUR **(n=487)**	SAS **(n=476)**
**A/*A*	Normal	21.0	23.2	3.8	11.8	0.6	2.7
**A/*B*	Normal	5.0	6.0	4.7	4.0	2.2	3.3
**B/*B*	Normal	0.3	0	0.9	1.9	3.2	0.2
**0/*A*	Reduced	45.9	39.1	24.7	37.8	14.9	26.5
**0/*B*	Reduced	2.7	3.3	16.2	10.5	22.7	13.8
**0/*0*	None	24.4	28.5	49.4	33.4	56.5	53.2
**A/*AX2*	Increased	0.5	0	0	0.4	0	0.4
**B/*Ax2*	Increased	0.1	0	0.3	0	0	0
**Ax2/*Bx2*	Increased	0	0	0	0.2	0	0

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants
*Diplotype frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

Supplementary Table 10: The distribution of *GSTT1* diplotypes and associated GSTT1 enzyme activity phenotypes in SSA compared with various global populations

Diplotypes	Enzyme activity	Diplotype frequencies (%)					
		SSA **(n=541)**	African American/ Afro-Caribbean **(n=122)**	AMR **(n=338)**	EAS **(n=492)**	EUR **(n=497)**	SAS **(n=477)**
**0/*0*	None	42.7	31.1	24.0	47.0	17.7	15.9
**0/*A*	Reduced	45.1	54.1	50.6	43.1	45.1	52.2
**A/*A*	Normal	12.2	14.8	25.1	10.0	36.8	31.7
**A/*Ax2*	Increased	0.0	0.0	0.3	0.0	0.4	0.2

SSA – Sub-Saharan African participants; AMR – admixed American participants; EAS – East Asian participants; EUR – European participants; SAS – South Asian participants
*Diplotype frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes

Appendix A

Approved Research Protocol

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

[CANDIDATE'S SURNAME: [Please print]	Malinga	FIRST NAME/S:	Thandeka Vuyiswa Bongiwe (V.B)
STUDENT NUMBER:	2152565		

CURRENT QUALIFICATIONS: BSc Human Physiology, Genetics and Psychology (Distinction)
BHSc Human Genetics

TEL: N/A	CELL: 079 428 9336	E-MAIL: 2152565@students.wits.ac.za	FAX: N/A

DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED: MSc (Med) Genomic Medicine

PART-TIME OR FULL-TIME: Full time

FIRST REGISTERED FOR THIS DEGREE:	TERM : 2023	YEAR: 1

DEPARTMENT: Human Genetics

TITLE OF PROPOSED RESEARCH: Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations

CANDIDATE'S SIGNATURE:	DATE: 25 April 2023
SUPERVISOR 1 (NAME & SURNAME): Dr David Twesigomwe	50% Supervision

SUPERVISOR'S QUALIFICATIONS: PhD

SUPERVISOR'S DEPARTMENT: Sydney Brenner Institute for Molecular Bioscience

SUPERVISOR'S ADDRESS / TEL / E-MAIL: david.twesigomwe@wits.ac.za

SUPERVISOR 2 (NAME & SURNAME): Dr Houcemeddine Othman	50% Supervision

SUPERVISOR'S QUALIFICATIONS: PhD

SUPERVISOR'S ADDRESS / TEL / E-MAIL: houcemeddine.othman@wits.ac.za

SYNOPSIS OF RESEARCH:

Introduction: Tuberculosis (TB) is an infectious disease affecting many African countries. Although TB is treatable, anti-TB drug resistance and adverse drugs reactions (ADRs) such as hepatotoxicity are a major health problem. These ADRs can partly be attributed to variants in pharmacogenes involved in anti-TB metabolism. The distribution of star alleles (haplotypes) that influence anti-TB drug metabolism, is unknown in many African ethnolinguistic groups. This presents challenges and limitations in implementing genotype-guided therapy in African clinical settings for purposes of anti-TB drug dosage optimisation with a view toward promoting safe and efficacious TB treatment. Therefore, characterising star alleles contributing to altered metabolism of anti-TB drugs among understudied African populations would be beneficial in informing pharmacogenetic testing and improving TB treatment.

Aim: To characterise the distribution of star alleles in genes that are involved in anti-TB drug metabolism, namely *CYP2E1*, *NAT1*, *NAT2*, *GSTM1* and *GSTT1*, across diverse African populations.

Methods: High-depth whole genome sequence datasets (n=802) representative of Sub-Saharan African populations included in five recent small- to large-scale human genomic studies will be analysed. The StellarPGx pipeline will be used to call known star alleles in *CYP2E1, NAT1, NAT2, GSTM1 and GSTT1,* and to identify potential novel star alleles. Novel star allele-defining variants will be annotated using the Ensembl Variant Effect Predictor (VEP). The Fisher's exact test will be used to assess the statistical significance of any observed differences in the distribution of various star alleles among the different African populations.

Expected outcomes: This study will provide insights into the distribution of common, rare and possible novel haplotypes that may account for differences in anti-TB interindividual drug response across Africa. This may enhance precision medicine in African populations, with a high TB burden, to reduce the incidence of ADRs to anti-TB drugs while maintaining the efficacy.

WITS ETHICS NOT REQUIRED:	Yes No	IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE:
WITS ETHICS PENDING:	Yes No	
WITS ETHICS APPROVED:	Yes No	
(circle appropriate symbol)*		
***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:	David Twesigomwe (Apr 25, 2023 19:00 GMT+2)	Houcemeddine Othman (Apr 25, 2023 19:07 GMT+1)
SIGNATURE PG OFFICE STAFF ……..	REGISTERED YES….. NO…..	STAMP

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Characterisation of the genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations

Research protocol

26/04/2023

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

MSc (Med) Genomic Medicine

By

Thandeka V.B. Malinga
Student number: 2152565

Supervisor 1: Dr David Twesigomwe
Supervisor 2: Dr Houcemeddine Othman

NATIONAL HEALTH LABORATORY SERVICE

SDIMB SYDNEY BRENNER INSTITUTE FOR MOLECULAR BIOSCIENCE

Table of Contents

1 Introduction 1
1.1 TB treatment and adverse drug reactions 1
1.2 Anti-TB drug metabolism and pharmacogenomics 1
1.2.1 N-acetyltransferase (NAT) 3
1.2.1.1 NAT2 3
1.2.1.2 NAT1 4
1.2.2 Cytochrome P450 2E1 (CYP2E1) 4
1.2.3 Glutathione S-Transferases (GST) 5
1.3 Study rationale 5
1.4 Aim 6
1.5 Objectives 6
2 Methodology 7
2.1 Study design 7
2.2 Data sources 7
2.3 Bioinformatics analysis 8
2.4 Variant annotation 9
2.5 Statistical analysis 9
3 Ethics 9
4 Timeline 10
5 Funding 10
6 Limitations 10
7 References 11

List of figures

Figure 1: Isoniazid (INH) metabolism in the liver 2
Figure 2: Overview of the genomic location of *NAT* genes and examples of key star alleles ...3
Figure 3: Summary of the Bioinformatics analysis workflow for the study 8

List of Tables

Table 1: WGS data of various African countries to be analysed with their respective data sources.....7
Table 2: Gantt chart.....10

1 Introduction

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* and is a major health burden in Africa. Moreover, in 2022, the World Health Organization (WHO) reported 17 African high TB burden countries, including Angola, Ethiopia, Namibia, and South Africa (WHO, 2022). Although TB is treatable, treatment regimens can be impacted by adverse drug reactions (ADRs) and drug-resistant strains. In addition, TB-HIV coinfection is common in African countries (WHO, 2022), which further complicates TB treatment efficacy due to drug-drug interactions (Manosuthi et al., 2013).

1.1 TB treatment and adverse drug reactions

First-line anti-TB drugs include isoniazid (INH), rifampin, pyrazinamide, and ethambutol. Second-line anti-TB drugs include fluoroquinolones, bedaquiline, and p-aminosalicylic acid (PAS). Combining first- and second-line anti-TB drugs aids in the treatment of TB strains resistant to first-line anti-TB drugs (ie. multidrug-resistant TB (MDR-TB)) (WHO, 2020).

Interindividual responses to anti-TB drugs vary across different populations and are attributable to environmental and host genetic factors. In particular, hepatotoxicity is a common and often fatal ADR to anti-TB drugs. Other ADRs include gastrointestinal, psychiatric, and vascular disorders (Bouazzi et al., 2016). These reactions may result in drug discontinuation (Pietersen et al., 2023), and an increased risk of MDR-TB. As genetic factors contribute to anti-TB drug response variability (Haas et al., 2022), genotype-guided treatment for individuals with TB may aid in reducing the risk of ADRs to anti-TB drugs and increase treatment efficacy (Azuma et al., 2013) in high TB burden African countries.

1.2 Anti-TB drug metabolism and pharmacogenomics

The enzymes NAT2, CYP2E1, GSTM1, and GSTT1, metabolise anti-TB drugs, such as INH (Figure 1) (Sannie et al., 2022), whereas NAT1 metabolises PAS (Cummins et al., 1997). Genetic variations in the respective enzyme-encoding genes have been associated with ADRs and drug response variability across various pharmacogenetic studies. For example, Jaramillo-Valverde et al (2022) showed that a combination of NAT2 fast acetylation, *GSTM1* non-null and *GSTT1* null genotypes, and rapid metabolism by CYP2E1 is associated with hepatotoxicity in individuals with TB who were treated with first-line anti-TB drugs. However, clinical variant annotations for newer second-line drugs such as bedaquiline are yet to be fully established.

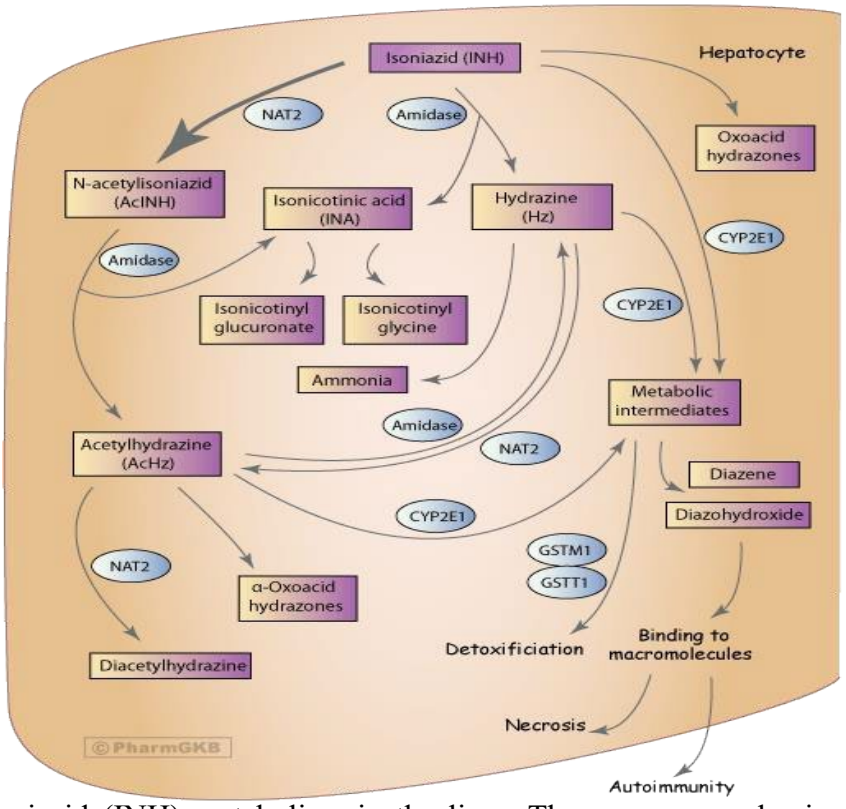


Figure 1: Isoniazid (INH) metabolism in the liver. The process can begin with either acetylation, hydrolysis of an amide group or oxidation by NAT2, amidase or CYP2E1 enzymes, respectively. The INH metabolites are detoxified by GSTM1 and GSTT1. Image credit: PharmGKB (https://www.pharmgkb.org/pathway/PA166151813) shared under Creative Commons BY-SA 4.0 license.

Haplotypes in *NAT2*, *NAT1*, *CYP2E1*, *GSTM1,* and *GSTT1* are represented via the star (*) allele nomenclature (Gaedigk et al., 2021). A star allele differing from the reference haplotype is typically defined by various combination(s) of missense, frameshift, stop gain/loss, splice site, and/or structural variants (see Figure 2 of key *NAT2* star alleles illustrating this notation). The definitions of star alleles in key pharmacogenes are mainly curated by the Pharmacogene Variation Consortium (PharmVar) to promote uniformity across research and clinical settings and support downstream genotype-phenotype translation efforts (Gaedigk et al., 2021).

In addition to the PharmVar resources, the Clinical Pharmacogenetics Implementation Consortium (CPIC) and the Pharmacogenomics Knowledge Base (PharmGKB) assess the evidence associating a drug-related response to specific genetic variation, thus aiding in pharmacogenetic testing (Whirl-Carrillo et al., 2021). According to PharmGKB's curation of existing pharmacogenetic studies, the star alleles in pharmacogenes associated with anti-TB drug metabolism have varying levels of clinical annotations. Specifically, *NAT2* is a Tier 1

very important pharmacogene (VIP) owing to numerous reports of *NAT2*-INH clinical variant annotations, whereas *CYP2E1* and *GSTT1* are considered Tier 2 VIPs (https://www.pharmgkb.org/vips). However, African populations are underrepresented in PharmGKB's curated studies. Therefore, further studies are needed to investigate function-altering star alleles that may impact anti-TB drug metabolism in African populations.

1.2.1 N-acetyltransferase (NAT)

N-acetyltransferases are phase II metabolising enzymes that aid in detoxifying drugs and carcinogens by acetylation (Liu et al., 2007). They are encoded by *NAT1* and *NAT2*, located on chromosome region 8p22 (Figure 2a). Acetylation activity by NAT1 and NAT2 ranges from rapid to slow acetylation largely due to genetic variation (Mcdonagh et al., 2014).

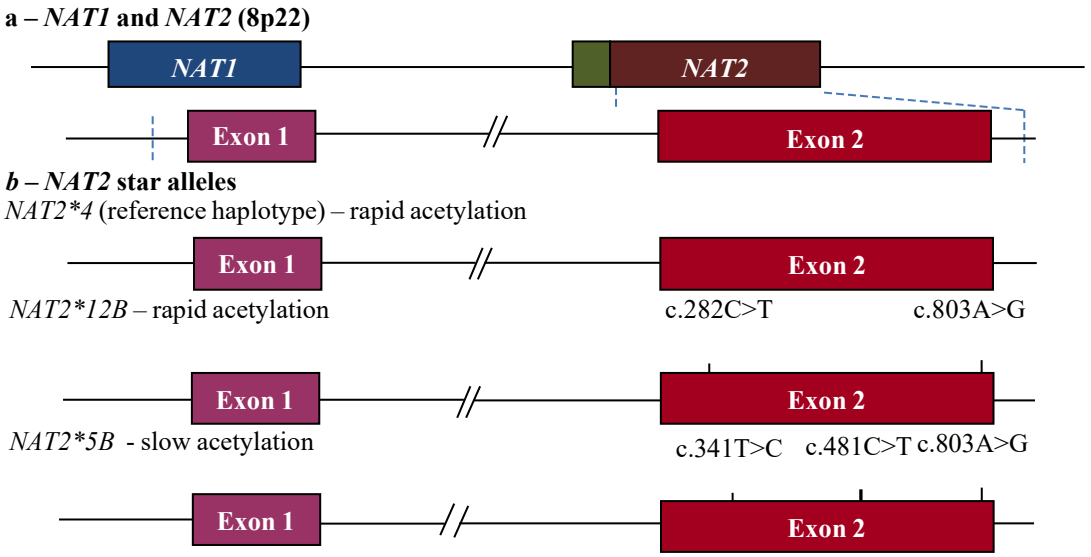


Figure 2: Overview of the genomic location of *NAT* genes and examples of key star alleles (a) *NAT1* and *NAT2* on chromosome 8. (b) Examples of *NAT2* star alleles and the corresponding allele-defining variants.

1.2.1.1 NAT2

NAT2 acetylates first-line anti-TB drugs, thus forming non-toxic metabolites (Bose et al., 2011). The encoding *NAT2* comprises two exons, one of which (exon 1) is untranslated. *NAT2* is highly polymorphic, with more than 20 known star alleles present, that are defined by variants in exon 2 (Figure 2b). While *NAT2* star alleles, such as **4* (rapid acetylation), have known functional impacts, many *NAT2* star alleles have unknown acetylation effects (http://nat.mbg.duth.gr/).

In African populations, *NAT2* star alleles associated with slow acetylation are more prevalent than those associated with rapid acetylation based on current literature. Specifically, *NAT2*5B*

(Figure 2b) has a frequency between 22-51% among South African, Republic of Congo, and Moroccan populations (Sabbagh et al., 2008). Given the significant role of NAT2 in anti-TB response and the phenotype variability associated with *NAT2* star alleles, continuous characterisation of known as well as novel *NAT2* star alleles using high-throughput African genomic datasets can be beneficial in improving clinical pharmacogenetic testing and, in turn, anti-TB drug response, especially in underrepresented populations.

1.2.1.2 NAT1

The second-line anti-TB drug PAS is metabolised by NAT1 (Cummins et al., 1997). As with *NAT2* star alleles, several known *NAT1* star alleles have unknown acetylation effects (http://nat.mbg.duth.gr/). Notably, *NAT1*3*, **14*, **15,* and **16* results in slow acetylation of PAS (Sy et al., 2015). *NAT1* star alleles have been well-characterised in various global populations (Walker et al., 2009); however, *NAT1* star allele distribution data among African populations is scarce. Loktionov et al (2002) found that the *NAT1*10* frequency (rapid acetylation), was higher in the Black South African population (50.5%) than in the European population (18.7%). In contrast, slow acetylation *NAT1* star alleles were either absent or less frequent in the Black South African population, apart from *NAT1*4*, which was common (48.5%) (Loktionov et al., 2002). Continuous efforts to identify *NAT1* star alleles using African datasets would provide insights to PAS drug responses and may enhance MDR-TB treatment.

1.2.2 Cytochrome P450 2E1 (CYP2E1)

The Cytochrome P450 (CYP) superfamily consists of phase I enzymes that metabolise drugs into their active metabolites (Zanger & Schwab, 2013). In particular, CYP2E1 (mainly expressed in the liver) is a significant enzyme involved in INH metabolism (Klein et al., 2016). *CYP2E1*, located on chromosome region 10q26.3, is highly polymorphic — consisting of star alleles defined by both single nucleotide variants, and structural variants such as deletions, insertions, and rearrangements (Martis et al., 2013). This should be considered when determining the distribution of *CYP2E1* star alleles and choosing appropriate techniques for pharmacogenetic testing of *CYP2E1*.

The wild-type star allele, *CYP2E1*1A*, results in normal CYP2E1 activity and is a commonly reported star allele in African countries, including Moroccan and South African populations (Guaoua et al., 2014; Heathfield et al., 2014). However, the distribution of function-altering alleles, such as *CYP2E1*5B* and *CYP2E1*1C,* has not been well studied in African countries,

despite their association with hepatotoxicity in patients with TB (Huang et al., 2003; Santos et al., 2013). Evidence associating *CYP2E1*1A, *5A*, and **5B* with hepatotoxicity has been rated as level 3 (low) by CPIC and PharmGKB (https://www.pharmgkb.org/). However, this may be due to inadequate studies on the genetic factors contributing to the ADRs of anti-TB drugs in African populations, which have a high TB burden compared to European populations.

Other CYP enzymes such as CYP3A4 and CYP3A5 are known to be involved in the metabolism of anti-TB drugs such as bedaquiline (Haas et al., 2022). However, much fewer clinical annotations have been reported regarding whether variations in these genes contribute to anti-TB ADRs. Therefore, these genes will not be included in the scope of this study.

1.2.3 Glutathione S-Transferases (GST)

The GST superfamily are phase II detoxification enzymes of xenobiotics and drug metabolites (Hayes et al., 2000). Two genes of the cytosolic family, *GSTM1* and *GSTT1*, are involved in INH metabolite detoxification (Figure 1). *GSTM1* is located on chromosome 1p13.3, while *GSTT1* is located on 22q11.23. Deletion of one or both genes result in reduced enzyme activity, thereby impeding substrate detoxification and leading to cellular damage (Hayes et al., 2000).

The *GSTM1* and *GSTT1* deletion (null) genotypes are frequent among African and non-African populations. In a Kassogue et al (2020) study, the frequencies of the *GSTM1* null (24.3%) and *GSTT1* null (41.3%) in a Malian population were comparable to Namibian and Gambian populations. However, in Moroccan and Algerian populations, the *GSTM1* null frequency was higher than the *GSTT1* null frequency (Kassogue et al., 2020). Due to their clinical impact, it would be advantageous to determine the *GSTM1* and *GSTT1* null deletion frequencies in African populations to inform precision medicine strategies, thereby possibly reducing ADRs that arise due to reduced detoxification of INH metabolites.

1.3 Study rationale

Star alleles in pharmacogenes that mediate the metabolism and detoxification of anti-TB drugs have been well characterised in European (Loktionov et al., 2002) and other global populations (Jaramillo-Valverde et al., 2022). Although these efforts have been replicated in some African populations, such as Morocco (Guaoua et al., 2014), Tanzania (Masiphephethu et al., 2022), Nigeria (Ebeshi et al., 2011), and South Africa (Heathfield et al., 2014), majority of the African ethnolinguistic groups remain vastly underrepresented. It is important to comprehensively

characterise common, rare, and possible novel star alleles that could impact metabolism of anti-TB drugs across Africa to build an evidence base for clinical pharmacogenomics (PGx) implementation. This is especially important in African countries with a high TB burden, such as South Africa, Sierra Leone, Namibia, and Kenya (WHO, 2022), and is crucial in addressing the burden of ADRs attributed to variations in pharmacogenes influencing anti-TB metabolism.

The increasing availability of high coverage whole genome sequencing data from research initiatives such as the Human Heredity & Health in Africa (H3Africa) projects (Choudhury et al., 2020) and the 1000 Genomes Project (Auton et al., 2015) provides an opportunity for comprehensive PGx research across previously understudied African populations. In addition, the development of bioinformatics pipelines, such as StellarPGx (Twesigomwe et al., 2021), can facilitate scalable star allele analysis based on these high-throughput African genomes. Specifically, using datasets consisting of underrepresented African populations can allow for the identification of novel African-ancestry function-altering star alleles in pharmacogenes involved in the metabolism of anti-TB drugs. This would aid in precision medicine implementation in Africa by highlighting star alleles that should be included in pharmacogenetic testing. Furthermore, such interventions would promote the accurate determination of anti-TB drug response phenotypes and inform appropriate drug dosage optimisation, with a view toward mitigating the risk of ADRs such as hepatotoxicity.

1.4 Aim

The aim of this study is to characterise the distribution of star alleles in genes that are involved in anti-TB drug metabolism, namely *CYP2E1*, *NAT1*, *NAT2, GSTM1* and *GSTT1*, across diverse African populations.

1.5 Objectives

1. To characterise the distribution of previously known star alleles in *CYP2E1*, *GSTM1*, *GSTT1*, *NAT1* and *NAT2* across different African populations.
2. To identify novel star alleles in *CYP2E1, GSTM1, GSTT1, NAT1* and *NAT2* across African populations.
3. To annotate and predict the functional effect of variants defining novel star alleles relevant to anti-TB pharmacogenomics across African populations.

2 Methodology

2.1 Study design

This is a comparative study investigating the star allele distribution in pharmacogenes involved in anti-TB drug metabolism between different Sub-Saharan African populations using secondary whole genome sequence (WGS) datasets.

2.2 Data sources

High-coverage WGS data of 802 African individuals will be analysed from the five data sources as indicated in Table 1. This includes publicly available genomes (30x) from 504 African participants in the 1000 Genomes Project (Byrska-Bishop et al., 2022), and public WGS data (43x) of 39 African participants from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016). In addition, the study will include genomes (40x read depth) of 100 south-eastern Bantu in South Africa, 33 Burkina Faso and 26 Ghana participants from the Africa Wits-INDEPTH partnership for Genomics (AWI-Gen) Research study (Ramsay et al., 2016), WGS (~30x) of 15 South African participants from the South African Human Genome Programme (SAHGP) (Choudhury et al., 2017), and WGS data (>30x read depth) of 85 participants from the Cell Biology Research Lab affiliated with the National Institute of Communicable Diseases (NICD)/ University of the Witwatersrand (Wits).

Table 1: WGS data of various African countries to be analysed with their respective data sources

African countries/Populations	**Primary project**	**n**
Public datasets		
Gambian in Western Division, Mandinka (GWD)	1000 Genome Project	113
Luhya in Webuye, Kenya (LWK)	1000 Genome Project	99
Esan in Nigeria (ESN)	1000 Genome Project	99
Yoruba in Ibadan (YRI)	1000 Genome Project	108
Mende in Sierra Leone (MSL)	1000 Genome Project	85
Various African ethnolinguistic groups	SGDP	39
African genomes from Africa-based research projects/collaborations		
South-eastern Bantu (SEB) in South Africa	AWI-Gen	100
Burkina Faso (BFA)	AWI-Gen	33
Ghana (GHA)	AWI-Gen	26
South Africa	SAHGP	15
South Africa	CBRL(NICD, WITS)	85

SGDP - Simons Genome Diversity Project, AWI-Gen - Africa Wits-INDEPTH partnership for Genomics, SAHGP - South African Human Genome Project, CBRL (NICD, WITS) - Cell Biology Research Lab, National Institute of Communicable Diseases, University of the Witwatersrand

2.3 Bioinformatics analysis

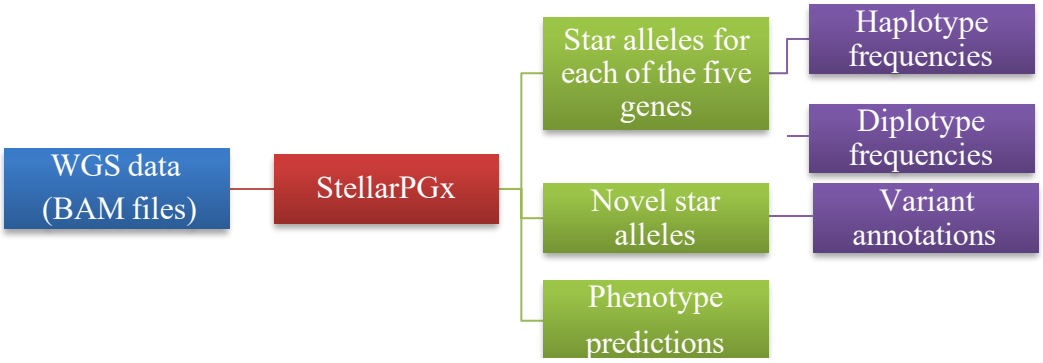


Figure 3: Summary of the Bioinformatics analysis workflow for the study. BAM, Binary Alignment Map.

StellarPGx (Twesigomwe et al., 2021), a Nextflow-based pipeline, will be used to call star alleles in *NAT1, NAT2, CYP2E1, GSTM1* and *GSTT1* from the five WGS datasets (Figure 3). This multi-step bioinformatics pipeline uses a BAM file, consisting of sequencing reads aligned to a reference genome, as an input. StellarPGx uses GraphTyper to re-align the reads to a variant-aware reference genome (ie. pangenome) (Eggertsson et al., 2017), consisting of all known variants in the pharmacogenes of interest, and subsequently to call variants. This method of variant calling is favoured, as the occurrence of reference allele bias or read misalignment to a reference genome is minimal. The GraphTyper output consists of variant calls and allele depth information stored in the variant call file (VCF) format.

For star allele assignment, StellarPGx compares allele-defining SNVs from the VCF to known star alleles present in the StellarPGx database. The database consists of precomputed diplotype combinations based mainly on star allele definitions curated by the PharmVar. In addition, the database has been updated to include known star allele definitions of *NAT1* and *NAT2* obtained from the Database of NATs (http://nat.mbg.duth.gr/). A diplotype is determined by matching the combination of core SNVs (typically nonsynonymous variants) in a sample to a diplotype in the database. In some cases, the combination of core SNVs from the VCF matches more than one diplotype in the database. To resolve this, suballele variants (non-function-altering variants in a haplotype) are considered to determine the appropriate candidate diplotype.

To call star alleles that consist of copy number variants in genes such as *CYP2E1*, the depth of coverage of the target gene is obtained using SAMtools *depth* (Li et al., 2009), and is compared to the read depth of the control loci, vitamin D receptor (*VDR)* and epidermal growth factor receptor (*EGFR*), to determine the copy number changes in key gene regions. *VDR* and *EGFR* are chosen as control regions (copy number=2) as they are relatively large genes with

fundamental biological functions and therefore, less likely to be wholly affected by CNVs. This is also supported by data from the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). This information is then combined with SNV allele depth information in the VCF files to ascertain the nature of the duplications, deletions and other structural rearrangements.

For samples with novel star alleles or suballeles, StellarPGx provides a flag in the output, and in addition it displays the core variants, and the possible background diplotype. This narrows down the number of samples which require additional manual inspection (e.g. visualisation on the Integrative Genomics Viewer) of novel star allele-defining variants.

2.4 Variant annotation

Annotation of variants defining novel star alleles will be performed using Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016). *In-silico* prediction tools including CADD, LRT, MutationAssessor, PROVEAN, Polyphen-2, SIFT, and VEST4, present as VEP-plugins, will be used (first separately) to predict the functional impact of these variants. A consensus functional prediction will be obtained from these tools based on parameters recommended by Zhou et al. (2019) with regards to pharmacogenetic variant impact assessment.

2.5 Statistical analysis

Star allele and diplotype frequencies will be summarised using percentages. In addition, Hardy-Weinberg Equilibrium (HWE) will be determined using the genetics package in R programming (https://www.r-project.org/), to detect deviations from expected frequencies. The Fisher's exact test will be used to assess the statistical significance of any observed frequency differences across populations.

3 Ethics

This study will involve an analysis of secondary WGS data. No contact will be made with participants from whom the WGS data was generated, and no new sequence data will be generated in the scope of this study. The data is already de-identified and will be analysed on the ZA-Wits-Core computing cluster (PI: Prof Scott Hazelhurst).

This sub-study will fall under the parent study, Approaches of GenOmics for Response to drugs in Africa-TB-Malaria (AGORA-TM). Ethics approval has been granted for the AGORA-TM

(ethics clearance certificate number: M220738). A sub-study ethics application for use of the SAHGP/AWI-Gen datasets (PI: Prof Michéle Ramsay), and the NICD datasets (PI: Prof Caroline Tiemessen) in this MSc research project is under consideration by the Wits Human Research Ethics Committee. Ethics approval will not be required for the 1000 Genomes Project or SGDP data, as these datasets are publicly available.

4 Timeline

Table 2: Gantt chart

	Feb	Mar	Apr	May	Jun	Jul	Aug	Oct	Nov
Literature review									
Ethics application									
Writing and presentation of research protocol									
Ethics approval									
Star allele calling and variant annotation									
Statistical analysis									
Research report writing									
Submission and presentation of research report									

5 Funding

This research will be funded by the Genomic Research Approach for Diversity and Optimising Therapeutics (GRADIENT) project, a partnership between GlaxoSmithKline, Novartis, and the South African Medical Research Council as part of the AGORA-TM project (PI: Dr Houcemeddine Othman). In addition, the student will be supported by the Wits Postgraduate Merit Award and the AGORA-TM project.

No budget is included with this proposal as (1) there are no data generation costs (2) the computational resources needed for the study are accessible via the ZA-Wits-Core Cluster (PI: Prof Scott Hazelhurst) (3) all bioinformatics tools to be used are open-source.

6 Limitations

The study population consists predominantly of sub-Saharan African populations, and in particular Niger-Congo language family ethnolinguistic groups. Therefore, star alleles and pharmacogenomic insights from other underrepresented African ethnolinguistic groups will not be investigated.

Phenotype data from the primary projects, that generated the genomic data, are not available. Therefore, correlation of pharmacogene star allele information to anti-TB drug response cannot

be made. In addition, functional studies to determine the CPIC clinical function of novel and/or rare star alleles are out of the scope of this research study. Functional studies would be beneficial in confirming any deleterious predictions made by the *in-silico* prediction tools.

The sample sizes of the Sub-Saharan African population groups in this study are relatively small. As a result, the star allele frequencies in these populations may be inaccurate and/or statistical analysis for these populations may be difficult.

7 References

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., et al. 2015. A global reference for human genetic variation. *Nature*. 526(7571):68–74.

Azuma, J., Ohno, M., Kubota, R., Yokota, S., Nagai, T., Tsuyuguchi, K., et al. 2013. NAT2 genotype guided regimen reduces isoniazid-induced liver injury and early treatment failure in the 6-month four-drug standard treatment of tuberculosis: A randomized controlled trial for pharmacogenetics-based therapy. *European Journal of Clinical Pharmacology*. 69(5):1091–1101.

Bouazzi, O.E., Hammi, S., Bourkadi, J.E., Tebaa, A., Tanani, D.S., Soulaymani-Bencheikh, R., et al. 2016. First line anti-tuberculosis induced hepatotoxicity: incidence and risk factors. *PanAfrican Medical Journal*. 25(167):1-6.

Bose, P.D., Sarma, M.P., Medhi, S., Das, B.C., Husain, S.A. & Kar, P. 2011. Role of polymorphic N-acetyl transferase2 and cytochrome P4502E1 gene in antituberculosis treatment-induced hepatitis. *Journal of Gastroenterology and Hepatology*. 26(2):312–318.

Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 185(18):3426–3440.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., et al. 2017. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*. 8(1):1–12.

Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensellak, T., et al. 2020. High-depth African genomes inform human migration and health. *Nature*. 586(7831):741–748.

Cummins, C.L., O'Neil, W.M., Soo, E.C., Lloyd, D.K. & Wainer, I.W. 1997. Determination of p-aninosalicylic acid and its N-acetylated metabolite in human urine by capillary zone electrophoresis as a measure of in vivo N-acetyltransferase 1 activity. *Journal of Chromatography B: Biomedical Sciences and Applications*. 697(1–2):283–288.

Ebeshi, B.U., Bolaji, O.O. & Masimirembwa, C.M. 2011. Arylamine N-acetyltransferase 2 (NAT2) single nucleotide polymorphisms' frequencies in Nigerian populations. *African Journal of Pharmacy and Pharmacology Research*. 1(1):1–6.

Eggertsson, H.P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., et al. 2017. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*. 49(11):1654–1660.

Gaedigk, A., Casey, S.T., Whirl-Carrillo, M., Miller, N.A. & Klein, T.E. 2021. Pharmacogene Variation Consortium: A Global Resource and Repository for Pharmacogene Variation. *Clinical Pharmacology & Therapeutics*. 110(3):542–545.

Guaoua, S., Ratbi, I., Laarabi, F.Z., Elalaoui, S.C., Jaouad, I.C., Barkat, A. & Sefiani, A. 2014. Distribution of allelic and genotypic frequencies of NAT2 and CYP2E1 variants in Moroccan population. *BMC Genetics*. 15(1): 156.

Haas, D.W., Abdelwahab, M.T., Van Beek, S.W., Baker, P., Maartens, G., Bradford, Y., et al. 2022. Pharmacogenetics of Between-Individual Variability in Plasma Clearance of Bedaquiline and Clofazimine in South Africa. *The Journal of Infectious Diseases*. 226(1):147–156.

Hayes, J.D., Strange, R.C. & Hayes, J.D. 2000. Glutathione S-Transferase Polymorphisms and Their Biological Consequences. *Pharmacology*. 61(3):154–166.

Heathfield, L.J., Dalvie, S., Kalideen, K. & Dandara, C. 2014. Novel CYP2E1 haplotype identified in a South African cohort. *South African Journal of Science*. 110(9–10):1–5.

Huang, Y.S., Chern, H.D, Su, W.J., Wu, J.C., Chang, S.C., Chiang, C.H., Chang, F.Y. & Lee, S.D. 2003. Cytochrome P450 2E1 genotype and the susceptibility to antituberculosis drug-induced hepatitis. *Hepatology*. 37(4):924–930.

Jaramillo-Valverde, L., Levano, K.S., Tarazona, D.D., Vasquez-Dominguez, A., Toledo-Nauto, A., Capristano, S., et al. 2022. GSTT1/GSTM1 Genotype and Anti-Tuberculosis Drug-Induced Hepatotoxicity in Peruvian Patients. *International Journal of Molecular Sciences*. 23(19):11028.

Kassogue, Y., Diakite, B., Kassogue, O., Konate, I., Tamboura, K., Diarra, Z., et al. 2020. Genetic polymorphism of drug metabolism enzymes (GSTM1, GSTT1 and GSTP1) in the healthy Malian population. *Molecular Biology Reports*. 47(1):393–400.

Klein, D.J., Boukouvala, S., McDonagh, E.M., Shuldiner, S.R., Laurieri, N., Thorn, C.F., et al. 2016. PharmGKB Summary: Isoniazid Pathway, Pharmacokinetics (PK). *Pharmacogenetics and Genomics*. 26(9):436–444.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25(16):2078–2079.

Liu, L., Von Vett, A., Zhang, N., Walters, K.J., Wagner, C.R. & Hanna, P.E. 2007. Arylamine N-acetyltransferases: Characterization of the substrate specificities and molecular interactions of environmental arylamines with human NAT1 and NAT2. *Chemical Research in Toxicology*. 20(9):1300–1308.

Loktionov, A., Moore, W., Spencer, S.P., Vorster, H., Nell, T., O'neill, I.K., Bingham, S.A. & Cummings, J.H. 2002. Differences in N-acetylation genotypes between Caucasians and Black South Africans: implications for cancer prevention. *Cancer Detection and Prevention*. 26(1):15–22.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 538(7624):201–206.

Manosuthi, W., Sukasem, C., Lueangniyomkul, A., Mankatitham, W., Thongyen, S., Nilkamhang, S., et al. 2013. Impact of pharmacogenetic markers of CYP2B6, clinical factors, and drug-drug interaction on efavirenz concentrations in HIV/tuberculosis-coinfected patients. *Antimicrobial Agents and Chemotherapy*. 57(2):1019–1024.

Martis, S., Mei, H., Vijzelaar, R., Edelmann, L., Desnick, R.J. & Scott, S.A. 2013. Multi-ethnic Cytochrome-P450 Copy Number Profiling: Novel Pharmacogenetic Alleles and Mechanism of Copy Number Variation Formation. *The Pharmacogenomics Journal*. 13(6):558–566.

Masiphephethu, M. V., Sariko, M., Walongo, T., Maro, A., Mduma, D., Gratz, J., et al. 2022. Pharmacogenetic testing for NAT2 genotypes in a Tanzanian population across the lifespan to guide future personalized isoniazid dosing. *Tuberculosis (Edinburgh, Scotland)*. 136: 102246.

Mcdonagh, E.M., Boukouvala, S., Aklillu, E., Hein, D.W., Altman, R.B. & Klein, T.E. 2014. PharmGKB Summary: Very Important Pharmacogene information for N-acetyltransferase 2. *Pharmacogenetic Genonomics*. 24(8):409–425.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., et al. 2016. The Ensembl Variant Effect Predictor. *Genome Biology*. 17(1):1–14.

Pietersen, E., Anderson, K., Cox, H., Dheda, K., Bian, A., Shepherd, B.E., et al. 2023. Variation in missed doses and reasons for discontinuation of anti-tuberculosis drugs during hospital treatment for drug-resistant tuberculosis in South Africa. *PloS One*. 18(2):e0281097.

Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., et al. 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:e20.

Sabbagh, A., Langaney, A., Darlu, P., Gérard, N., Krishnamoorthy, R. & Poloni, E.S. 2008. Worldwide distribution of NAT2 diversity: Implications for NAT2 evolutionary history. *BMC Genetics*. 9(1):1–14.

Santos, N.P., Callegari-Jacques, S.M., Dos Santos, A.K., Silva, C.A., Vallinoto, A.C., Fernandes, D.C., et al. 2013. N-acetyl transferase 2 and cytochrome P450 2E1 genes and isoniazid-induced hepatotoxicity in Brazilian patients. *International Journal of Tuberculosis and Lung Disease*. 17(4):499–504.

Sanni, S., Sina, H., & Baba-Moussa, L. (2022). Genetic Polymorphisms and Toxicities of First-Line Antituberculosis Drugs: Systematic Review of the Literature. *Journal of Tuberculosis Research*., *10*(3),: 124–145.

Sy, S.K., De Kock, L., Diacon, A.H., Werely, C.J., Xia, H., Rosenkranz, B., et al. 2015. N-acetyltransferase genotypes and the pharmacokinetics and tolerability of para-aminosalicylic acid in patients with drug-resistant pulmonary tuberculosis. *Antimicrobial Agents and Chemotherapy.* 59(7):4129–4138.

Twesigomwe, D., Drögemöller, B.I., Wright, G.E., Siddiqui, A., da Rocha, J., Lombard, Z. & Hazelhurst, S. 2021. StellarPGx: A Nextflow Pipeline for Calling Star Alleles in Cytochrome P450 Genes. *Clinical Pharmacology and Therapeutics*. 110(3):741–749.

Walker, K., Ginsberg, G., Hattis, D., Johns, D.O., Guyton, K.Z. & Sonawane, B. 2009. Genetic Polymorphism in N-Acetyltransferase (NAT): Population Distribution of NAT1 and NAT2 Activity. *Journal of Toxicology and Environmental Health, Part B*. 12(5–6):440–472.

Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C.F., Whaley, R. & Klein, T.E. 2021. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology and Therapeutics*. 110(3):563–572.

Zanger, U.M. & Schwab, M. 2013. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*. 138(1):103–141.

Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M., & Lauschke, V.M. 2019. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *The Pharmacogenomics Journal*. 19(2):115–126.

Internet references

World Health Organization. 2020. WHO consolidated guidelines on tuberculosis. Available: https://www.who.int/publications/i/item/9789240007048 [Accessed: 04.03.2023].

World Health Organization. 2022. Global Tuberculosis report 2022. Available: http://apps.who.int/bookorders [Accessed: 11.02.2023].

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: Ms T Malinga
School of Pathology
Division of Human Genetics
Medical School
University

E-mail: 2152565@students.wits.ac.za

CC: Supervisor: Drs T Twesigomwe and H Othman
<David.Twesigomwe@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: 2023/07/19

REF: R14/49

PROTOCOL NO: **M230687** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Characterization of genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across 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 Ms T Malinga

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M230687

NAME: (Principal Investigator)	Ms T Malinga
DEPARTMENT:	School of Pathology Division of Human Genetics Medical School University
PROJECT TITLE:	*Characterization of genetic variation in pharmacogenes involved in anti-tuberculosis drug metabolism across African populations*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M22/07/38
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Drs T Twesigomwe and H Othman
APPROVED BY:	Dr M Vorster, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	2023/07/19 **EXPIRY DATE:** 2028/07/18

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 **June** and therefore reports and re-certification will be due in the month of **June** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

Date

Appendix C

Turn-it-in Report

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Thandeka Vuyiswa Bongiwe Malinga (Student number: 2152565) am a student registered for the degree of MSc (Med) Genomic medicine in the academic year 2024.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: 26/02/2024

Final report_content for turnitin.docx

ORIGINALITY REPORT

13%	8%	11%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	David Twesigomwe, Britt I. Drögemöller, Galen E. B. Wright, Clement Adebamowo et al. " Characterisation of pharmacogenetic variation in African populations ", Clinical Pharmacology & Therapeutics, 2022 Publication	1%
2	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%
3	Ross P Booyse, David Twesigomwe, Scott Hazelhurst. " Characterization of pharmacogenetic variation in African populations and comparison with other global populations ", Pharmacogenomics, 2023 Publication	1%
4	zenodo.org Internet Source	1%
5	link.springer.com	

Appendix D

Author Guidelines for *Frontiers in Pharmacology*

Article Title

First Author[1], Second Author[2*], Third Author[1,2]

[1]Laboratory X, Institute X, Department X, Organization X, City X, State XX (only USA, Canada and Australia), Country

[2]Laboratory X, Institute X, Department X, Organization X, City X, State XX (only USA, Canada and Australia), Country

*** Correspondence:**
Corresponding Author
email@uni.edu

Keywords: keyword$_1$, keyword$_2$, keyword$_3$, keyword$_4$, keyword$_5$. (Min.5-Max. 8)

Abstract

For full guidelines please refer to Author Guidelines

As a primary goal, the abstract should render the general significance and conceptual advance of the work clearly accessible to a broad readership. References should not be cited in the abstract. Leave the Abstract empty if your article does not require one – please see the "Article types" on every Frontiers journal page for full details.

1 Introduction

For **Original Research Articles, Clinical Trial Articles**, and **Technology Reports** the introduction should be succinct, with no subheadings. For **Case Reports** the Introduction should include symptoms at presentation, physical exams and lab results.

2 Article types

For requirements for a specific article type please refer to the Article Types on any Frontiers journal page. Please also refer to Author Guidelines for further information on how to organize your manuscript in the required sections or their equivalents for your field[1].

3 Manuscript Formatting

3.1 Headings

[1] For Original Research articles, please note that the Material and Methods section can be placed in any of the following ways: before Results, before Discussion or after Discussion.

You may insert up to 5 heading levels into your manuscript as can be seen in "Styles" tab of this template. These formatting styles are meant as a guide, as long as the heading levels are clear, Frontiers style will be applied during typesetting.

3.2 Equations

The equations should be inserted in editable format from the equation editor.

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$$

3.3 Figures

Frontiers requires figures to be submitted individually, in the same order as they are referred to in the manuscript. Figures will then be automatically embedded at the bottom of the submitted manuscript. Kindly ensure that each table and figure is mentioned in the text and in numerical order. Figures must be of sufficient resolution for publication. Figures which are not according to the guidelines will cause substantial delay during the production process. Figure legends should be placed at the end of the manuscript. Please see here for full Figure guidelines

3.3.1 Permission to reuse and Copyright

Permission must be obtained for use of copyrighted material from other sources (including the web). Please note that it is compulsory to follow figure instructions.

3.4 Tables

Tables should be inserted at the end of the manuscript. Tables must be provided in an editable format e.g., Word, Excel. Tables provided as jpeg/tiff files will **not be accepted**. Please note that very large tables (covering several pages) cannot be included in the final PDF for reasons of space. **These tables will be published as Supplementary Material on the online article page at the time of acceptance. The author will be notified during the typesetting of the final article if this is the case.**

4 Nomenclature

4.1 Resource Identification Initiative

To take part in the Resource Identification Initiative, please use the corresponding catalog number and RRID in your current manuscript. For more information about the project and for steps on how to search for an RRID, please click here.

4.2 Life Science Identifiers

Life Science Identifiers (LSIDs) for ZOOBANK registered names or nomenclatural acts should be listed in the manuscript before the keywords with the following format:

urn:lsid:<Authority>:<Namespace>:<ObjectID>[:<Version>]

This is a provisional file, not the final typeset article

5 Additional Requirements

For additional requirements for specific article types and further information please refer to “Article types” on every Frontiers journal page.

6 Conflict of Interest

All financial, commercial or other relationships that might be perceived by the academic community as representing a potential conflict of interest must be disclosed. If no such relationship exists, authors will be asked to confirm the following statement:

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Author Contributions

The Author Contributions section is mandatory for all articles, including articles by sole authors. If an appropriate statement is not provided on submission, a standard one will be inserted during the production process. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work. Please see here for full authorship criteria.

8 Funding

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

9 Acknowledgments

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

10 Reference styles

The following formatting styles are meant as a guide, as long as the full citation is complete and clear, Frontiers referencing style will be applied during typesetting.

10.1 Science, Engineering and Humanities and Social Sciences references

For articles submitted in the domains of Science, Engineering or Humanities and Social Sciences please apply Author-Year system for in-text citations. For Humanities and Social Sciences articles please include page numbers in the in-text citations

For some examples please click here.

For more examples of citing other documents and general questions regarding reference style, please refer to the Chicago Manual of Style.

10.2 Health, Physics and Mathematics references

For articles submitted in the domain of Health or the journals Frontiers in Physics and Frontiers in Applied Mathematics and Statistics please apply the Vancouver system for in-text citations.

In-text citations should be numbered consecutively in order of appearance in the text – identified by Arabic numerals in the parenthesis [square parenthesis for Physics and Mathematics].

For some examples please click here.

For more examples of citing other documents and general questions regarding reference style, please refer to Citing Medicine.

11 Supplementary Material

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. Supplementary Material templates can be found in the Frontiers Word Templates file.

Please see the Supplementary Material section of the Author guidelines for details on the different file types accepted.

12 Data Availability Statement

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK]. Please see the "Availability of data" section of Materials and data policies in the Author guidelines for more details.

This is a provisional file, not the final typeset article