

Characterization of *NAT*, *GST*, and *CYP2E1* Genetic Variation in Sub-Saharan African Populations: Implications for Treatment of Tuberculosis and Other Diseases

Thandeka V.B. Malinga^{1,2} , Houcemeddine Othman^{1,3} , Maria Paximadis^{4,5} ,
Caroline T. Tiemessen⁵ , Michèle Ramsay¹ , Scott Hazelhurst^{1,6}  and David Twesigomwe^{1,*} 

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. In this study, we used StellarPGx to call variants and star alleles in *NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1*, from 1079 high-depth African whole genomes. We present the distribution of common, rare, and potential novel star alleles across various Sub-Saharan African (SSA) populations, in comparison with other global populations. *NAT1**10 (53.6%), *GSTT1**0 (65%), *GSTM1**0 (48%), and *NAT2**5 (17.5%) were among the

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

✓ Tuberculosis is a major health burden in Africa; however, its treatment is affected by variable interindividual drug responses which can include drug-induced liver injury. Genetic variation in *NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1* have been associated with differences in response to isoniazid (first-line anti-tuberculosis medication), and other medications for other treatments, for example, hydralazine (cardiovascular medication), and anti-cancer medications.

WHAT QUESTION DID THIS STUDY ADDRESS?

✓ The current distribution of variants and/or star alleles (haplotypes) in pharmacogenes involved in anti-TB drug metabolism is not well-characterized across African populations, thus limiting PK/PD and other functional pharmacogenetic studies, and development of comprehensive pharmacogenetic testing platforms. This study provides a comprehensive mapping of the common, rare, and novel genetic variation in *NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1*, among multiple SSA populations, and comparisons with global biogeographical groups.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

✓ This study elucidates the genetic diversity in *NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1* across African populations, including variants that have been previously associated

with adverse drug reactions such as anti-TB drug-induced hepatotoxicity or liver injury—especially in patients taking isoniazid. Differences in the distribution of predicted acetylation (*NAT1* and *NAT2*) and detoxification (*GSTM1* and *GSTT1*) phenotypes were observed in SSA compared with other global populations. Additionally, in keeping with the extensive genetic diversity among SSA populations, 47 possible novel star alleles were identified in *NAT1*, *NAT2*, *GSTM1*, and *GSTT1*, combined. Furthermore, we highlight estimated common and rare *GSTP1* and *CYP2E1* haplotypes across Africa, most of which comprised nonsynonymous variants not previously used to define *GSTP1* and *CYP2E1* star alleles.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

✓ This study adds to the growing efforts to inform pharmacogenetics testing and to enhance precision medicine in Africa, particularly in the context of TB treatment. The differences in allele distributions between SSA and other global populations indicate the need to critically assess the transferability of current pharmacogenetic-based dosing guidelines to SSA populations. In addition, the information on *NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1* common, rare and novel variants/star alleles among various SSA populations is critical in developing effective pharmacogenetic testing panels for use in Africa and for providing insights for future in vivo or in vitro functional pharmacogenetic studies.

predominant functionally relevant star alleles. Additionally, we predicted varying phenotype distributions for NAT1 and NAT2 (acetylation) and the glutathione-S-transferase (GST) enzymes (detoxification activity) between SSA and other global populations. Forty-seven potentially novel haplotypes were identified computationally across the genes. This study provides insight into the distribution of key variants and star alleles potentially relevant to anti-TB drug metabolism and other drugs prescribed across various African populations. The high number of potentially novel star alleles exemplifies the need for pharmacogenomics studies in the African context. Overall, our study provides a foundation for functional pharmacogenetic studies and potential implementation of pharmacogenetic testing in Africa to reduce the risk of ADRs related to treatment of TB and other diseases.

¹Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; ²Division of Human Genetics, National Health Laboratory Service, and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; ³Laboratory of Cytogenetics, Molecular Genetics and Reproductive Biology (LR03SP02), Farhat Hached University Hospital, Sousse, Tunisia; ⁴School of Molecular and Cell Biology, Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa; ⁵Centre for HIV and STIs, National Institute for Communicable Diseases, National Health Laboratory Services and Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; ⁶School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa. *Correspondence: David Twesigomwe (david.twesigomwe@wits.ac.za), Houcemeddine Othman (houcemeddine.othman@wits.ac.za)

Tuberculosis (TB) is a major public health concern in many developing countries worldwide and its treatment typically involves administering isoniazid (INH), rifampin, pyrazinamide, and/or ethambutol as first-line drugs.¹ Treatment of TB is complicated by drug resistance and—the focus of this paper—interindividual variability in drug responses.^{2,3} Anti-TB drug responses can range from subtherapeutic effects to adverse drug reactions (ADRs) such as anti-TB drug-induced hepatotoxicity or liver injury (AT-DILI), which has an incidence of up to 15% across Sub-Saharan Africa (SSA)^{4,5} and in global settings.⁶ This variability in treatment responses often prompts interruptions to treatment, in addition to the increased risk of drug-resistant strains emerging, thus leading to poor treatment outcomes and undermining TB eradication efforts.

AT-DILI and the overall variability in anti-TB drug response have been attributed to both genetic and non-genetic factors (e.g., age, comorbidities, drug–drug interactions, and lifestyle habits).⁷ Regarding the genetic component, variants in Cytochrome P450 2E1 (*CYP2E1*), N-acetyltransferase 2 (*NAT2*), and Glutathione-S-transferases (*GSTM1*, *GSTT1*, and *GSTP1*) have previously been associated with an increased AT-DILI risk, especially in individuals on INH.^{8,9} These genes may therefore be important to consider for genotype-guided TB treatment as this may mitigate the risk of ADRs and increase treatment efficacy. However, their pharmacogenetic diversity is understudied across African populations—many of which have a high TB burden¹—and could be a significant hindrance to precision medicine approaches.

Genetic variation in *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1* 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 has been used to denote these haplotypes. Besides 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). Variants in major star allele definitions may also be referred to as core variants, according to the Pharmacogene Variation (PharmVar) Consortium database (<https://www.pharmvar.org>). Based on the ever-expanding

pharmacogenetics body of knowledge over the last three decades, the Clinical Pharmacogenetics Implementation Consortium (CPIC) (<https://cpicpgx.org/>), in collaboration with other pharmacogenomics groups, has established genotype-guided drug dosing recommendations with a view toward reducing unfavorable 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 (e.g., *NAT2**5, *GSTM1**0, and *GSTT1**0) are known to be associated with an increased risk of AT-DILI.^{10,11}

NAT2 has recently been transitioned from <https://nat.mbg.duth.gr/> to PharmVar (<https://www.pharmvar.org/gene/NAT2>). However, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1* are yet to be considered for inclusion. PharmVar has currently cataloged 35 *NAT2* star alleles (<https://www.pharmvar.org/gene/NAT2>).

A major challenge in generating comprehensive evidence to support broader anti-TB pharmacogenetics studies is the scarcity in information regarding the distribution of star alleles potentially impacting the function of anti-TB drug metabolizing enzymes, especially across SSA, where the majority of the populations and/or ethnolinguistic groups are underrepresented in pharmacogenomics studies.¹² Therefore, this study aimed to leverage recently generated African whole genome sequencing (WGS) datasets to characterize the distribution of star alleles in *NAT2*, *GSTM1*, *GSTT1*, *GSTP1*, and *CYP2E1* across diverse African populations as these genes are key for the metabolic pathway of INH (Figure S1) and can also impact response to other anti-TB drugs (rifampin, pyrazinamide, ethambutol, and para-aminosalicylic acid). Understanding the landscape of common, rare and/or novel variants/star alleles in these genes could inform prioritization of variants for further *in vitro*, *in vivo*, and *in silico* functional studies pertinent for optimization of dosage regimens for INH and potentially other anti-TB drugs.

Importantly, *NAT2* (hydralazine), *GSTs* (cisplatin, nevirapine, imatinib, clozapine, doxorubicin, methotrexate, fluorouracil, and platinum compounds), and *CYP2E1* (cisplatin, risperidone, and cyclophosphamide) are also relevant for response to medicines for other diseases.^{13–15} Our analysis herein also includes *NAT1* given its genomic proximity to *NAT2*, and its importance in the metabolism of cotinine, thioguanine and mesalazine.

MATERIALS AND METHODS

Ethics

Ethical clearance for this study was granted by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Wits) 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 access all the non-public datasets used in this study was granted by the respective principal investigators concerned and/or the Human Heredity and Health in Africa (H3Africa) Database Access Committee, where applicable.

Study design

This was a comparative study entailing secondary analysis of WGS data to characterize the star allele distribution in pharmacogenes involved in anti-TB drug metabolism across various SSA populations, and in comparison, with global populations. Given the lack of current comprehensive star allele definitions for *CYP2E1*, we analyzed variation in this gene independent of the star allele annotations.

Study population and whole genome sequencing data sources

High-coverage WGS data (~30×) of 1079 African individuals were analyzed from five datasets (Table S1) and the participants were predominantly from Niger-Congo language family ethnolinguistic groups; the 1000 Genomes Project¹⁶ public dataset consisted of 504 genomes (30×) representative of five different SSA populations (Table S1). This included Gambian in Western Division, Mandinka (GWD; *n* = 113), Yoruba in Ibadan (YRI, *n* = 108), Esan in Nigeria (ESN, *n* = 99), Luhya in Webuye, Kenya (LWK, *n* = 99), and Mende in Sierra Leone (MSL, *n* = 85). Additional 1000 Genomes Project 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)¹⁷ included 30 genomes (43×) of participants from various SSA countries, including the Congo, Namibia, Kenya, Senegal, Nigeria, Sudan, and South Africa. Three samples had ambiguous country designations and were documented as Botswana/Namibia by the SGDP.

The H3Africa-Baylor dataset comprised 346 WGSs representative of various SSA populations including Benin (FNB, Fon in Benin; *n* = 50), Botswana (BOT, *n* = 47), Cameroon (CAM; *n* = 50), Mali (MAL, People from Mali, *n* = 50), Nigeria (BRN, the Berom of Nigeria, *n* = 49), Zambia (BSZ, Bantu speakers in Zambia; *n* = 41), Burkina Faso (BFA; *n* = 33), and Ghana (GHA; *n* = 26).

The Africa Wits-INDEPTH partnership for Genomic Studies (AWI-Gen) dataset comprised 100 genomes (~40×) from southeastern Bantu-speaking participants in South Africa (SA).¹⁸ The South African data also included 15 genomes (> 30×) obtained from the Southern African Human Genome Programme (SAHGP)¹⁹ and 84 genomes (> 30×) from the Cell Biology Research Lab affiliated with the National Institute of Communicable Diseases (NICD) and the University of the Witwatersrand.

All the binary alignment map (BAM) files comprised of reads aligned to GRCh38 and had already passed sample quality control from the primary studies highlighted above.

Star allele bioinformatics analysis

Star alleles in *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* were called using StellarPGx.²⁰ Individual WGS BAM files were used as input (Figure S2). Briefly, in StellarPGx v1.2.6, variants are called via GraphTyper,²¹

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. Variants with quality by depth (QD) < 10 and heterozygous variants with an allele balance for heterozygous calls (ABHet) < 0.25 or ABHet > 0.75 are filtered out by the pipeline as per conservative thresholds for hard filtering.²¹ 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 pre-computed diplotype combinations based on star allele definitions curated by PharmVar. For *NAT2*, we used the recently updated star allele nomenclature on the PharmVar database (<https://www.pharmvar.org/gene/NAT2>). However, since *CYP2E1*, *NAT1*, *GSTM1*, *GSTT1*, and *GSTP1* had not yet been included in the PharmVar database at the time of this study, we used *GSTM1* and *GSTT1* allele designations from published literature,^{22,23} and *NAT1* star allele definitions from the database of arylamine *N*-acetyltransferases (NATs) (<http://nat.mbg.duth.gr/>). For *CYP2E1* and *GSTP1*, we estimated and defined haplotypes present in the study populations using SHAPEIT (version 4)²⁴ as they lacked a fully curated star allele nomenclature at the time of the study. Previously published *GSTP1* and *CYP2E1* star allele definitions were referred to on a case-by-case basis where comparisons were appropriate.

For samples with potentially novel star alleles, StellarPGx provides a flag in the output, and in addition, it displays the core variants and the possible backbone alleles (typically *1 or other common known star alleles). This narrows down the number of samples that require additional manual inspection (e.g., visualization on the Integrative Genomics Viewer) of novel star allele defining variant haplotypes.

Phenotype prediction

The phenotype prediction was based on combining the associated predicted 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. and other studies, whereby three phenotype categories are possible, that is, NAT2 slow acetylator (presence of two star alleles associated with slow acetylation, e.g., *NAT2*5*), 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*).²⁵ A similar approach was adopted to infer NAT1 phenotypes. For *GSTM1* and *GSTT1*, derived phenotypes included normal function (two alleles with normal activity), reduced function (one allele with normal activity and one null allele), no function (two null alleles), and increased function (two increased activity alleles).

Variant annotation and functional prediction

Alleles defining novel haplotypes were annotated using the Ensembl Variant Effect Predictor (VEP).²⁶ Functional predictions of allele defining variants were inferred using the *in silico* VEP (Figure S2). A variant was considered deleterious if ≥ 3 *in silico* prediction tools predicted it to be deleterious. Stop-gained, frameshift, splice donor/acceptor variants were automatically considered deleterious. Additionally, DynaMut2²⁷ was used to predict their potential impact on the stability of the respective protein structures.

Statistical analysis

Hardy-Weinberg Equilibrium (HWE) for individual variants and star alleles was assessed using the genetics package (v1.3.8.1.3) in R programming (<https://www.r-project.org/>), with *P* < 0.05 indicating deviation

Table 1 Distribution of NAT1 and NAT2 star alleles across SSA populations in this study

Star alleles	Acetylation effect	Allele frequencies (%) ^a														
		All (n=1060)	GWD (n=112)	MSL (n=81)	MAL (n=50)	BFA (n=33)	GHA (n=26)	FNB (n=50)	BRN (n=49)	ESN (n=96)	YRI (n=106)	CAM (n=50)	BSZ (n=41)	BOT (n=47)	SA (n=194)	LWK (n=97)
NAT1																
*10	Rapid	53.6	58.3	59.3	56.0	65.1	61.5	54.0	61.2	51.0	47.2	57.6	34.0	53.8	54.1	45.4
*3	Slow	2.8	1.8	3.7	6.0	7.6	3.8	8.0	1.0	1.6	1.9	3.0	4.0	0	1.8	2.6
*4	Slow	42.0	37.7	36.4	37.0	27.3	32.7	36.0	35.7	46.4	50.0	36.4	42.0	44.1	42.6	50.5
*15	Slow	0.1	0.5	0	0	0	0	0	1.0	0	0	0	0	0	0.3	0
*27	Slow	1.2	0.5	0.6	0	0	1.9	2.0	0	1.0	0.9	3.0	1.0	2.3	1.3	1.5
*26B	Unknown	0.3	1.3	0	0	0	0	0	1.0	0	0	0	0	0	0	0
NAT2																
Allele frequencies (%) ^a																
Star alleles	Acetylation effect	All (n=1076)	GWD (n=112)	MSL (n=85)	MAL (n=50)	BFA (n=32)	GHA (n=26)	FNB (n=50)	BRN (n=49)	ESN (n=98)	YRI (n=108)	CAM (n=50)	BSZ (n=41)	BOT (n=47)	SA (n=199)	LWK (n=99)
NAT2																
*1	Rapid	17.1	11.9	19.4	9.0	20.0	11.5	14.1	13.3	14.2	15.3	14.1	12.2	27.7	27.0	10.1
*4	Rapid	15.2	14.2	21.2	19.0	12.3	11.5	19.2	15.3	15.2	22.7	7.1	12.2	6.4	13.8	14.6
*5	Slow	17.5	21.2	10.6	22.0	16.9	15.4	5.1	24.5	15.7	13.4	20.2	11.0	21.3	18.8	22.2
*6	Slow	12.9	9.3	15.3	10.0	10.8	17.3	14.1	17.3	16.8	11.1	14.1	17.1	14.9	11.0	11.6
*7	Slow	2.0	1.8	3.5	7.0	0	1.9	6.1	2.0	1.0	2.8	0	2.4	2.1	0.75	1.0
*14	Slow	10.1	14.1	7.1	9.0	12.3	13.5	18.2	13.3	12.7	9.7	12.1	12.2	8.5	6.5	7.6
*16	Slow	10.8	11.9	12.9	9.0	12.3	13.5	15.2	9.2	11.2	10.2	15.2	12.2	11.7	7.0	12.1
*27	Unknown	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0
*30	Slow	0.04	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0
*31	Unknown	0.4	0	0	0	0	0	0	0	0	0	0	0	1.1	1.0	1.5
*33	Unknown	0.3	0	0.6	0	0	0	0	0	0.5	0.5	0	0	0	1.0	0
*34	Unknown	5.2	4.4	1.8	5.0	3.1	5.8	6.1	4.1	5.1	3.7	12.1	9.8	4.3	3.5	10.1
*36	Unknown	1.5	0.4	0	1.0	0	1.9	0	0	0	0.5	0	3.7	1.1	4.3	4.0
*37	Unknown	0.1	0.9	0.6	0	0	0	0	0	0	0	0	0	0	0	0
*38	Unknown	1.1	0.4	0.6	4.0	1.5	0	1.0	0	0.5	0.5	0	0	1.1	2.5	0.5
*39	Unknown	2.5	3.1	4.1	4.0	7.7	1.9	0	0	3.0	3.7	4.0	4.9	0	1.3	1.0
*40	Slow	0.5	0	1.2	0	1.5	3.8	1.0	1.0	0.5	0.9	0	0	0	0	0
*41	Unknown	0.2	1.3	0	0	0	0	0	0	0	0.9	0	0	0	0	0
*42	Unknown	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
*43	Unknown	0.7	0.9	1.2	0	1.5	1.9	0	0	1.5	1.4	0	1.2	0	0.5	0

(Continued)

Table 1 (Continued)

Star alleles	Acetylation effect	Allele frequencies (%) ^a														
		All (n=1076)	GWD (n=112)	MSL (n=85)	MAL (n=50)	BFA (n=32)	GHA (n=26)	FNB (n=50)	BRN (n=49)	ESN (n=98)	YRI (n=108)	CAM (n=50)	BSZ (n=41)	BOT (n=47)	SA (n=199)	LWK (n=99)
*44	Unknown	1.0	4.0	0	1.0	0	0	0	0	1.5	0.9	0	0	0	0.5	1.5
*47	Unknown	0.4	0	0	0	0	0	0	0	0	1.4	1.0	0	0	0	1.5
*49	Slow	0.04	0	0	0	0	0	0	0	0	0	0	1.2	0	0	0
*50	Unknown	0.04	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0

BFA, Burkina Faso participants; BOT, Botswana participants; BRN, Berom of Nigeria; BSZ, Bantu speakers from Zambia; CAM, Cameroonian participants; ESN, Esan in Nigeria; FNB, Fon from Benin; GHA, Ghanaian participants; GWD, Gambian in Western Division Mandinka; LWK, Luhya in Webuye, Kenya; MAL, Malian participants; MSL, Mende in Sierra Leone; SA, South African participants; YRI, Yoruba in Ibadan (Nigeria).
^aAllele frequency estimates excluded individuals with unresolved diplotypes. Given the novel alleles inferred in the study (see Table 6), the NAT1*4 and NAT2*1 frequencies should be considered as is and not “100 - (total frequencies of other alleles)”. The acetylation effects indicated are based on those reported in literature (ref 29) and the database of NATs (<http://nat.mbg.duth.gr/>). Further information (e.g. allele definitions) was obtained from PharmVar (<https://www.pharmvar.org/>; last accessed date: October 17, 2024) and PharmGKB (<https://www.pharmgkb.org/>; last accessed date: October 17, 2024).

from HWE. Fisher’s exact test was used to assess the statistical significance of any observed frequency differences across the populations. *P* values < 0.05 were considered statistically significant after correcting for multiple testing using the Benjamini–Yekutieli method.²⁸

RESULTS
Summary of variants

A total of 2031 SNVs were identified across the six anti-TB metabolism pharmacogenes in the SSA populations analyzed in this study. The breakdown of the variant consequences is shown in Figure S3. We observed a higher variant count—32 variants per kilobase (kb)—in SSA compared with the other five global biogeographical groups (16–18 variants per kb). For the individual genes, *GSTP1* (~154 variants per kb) had the highest number of variants followed by *NAT1* (~38 variants per kb) in SSA.

NAT1 and NAT2 star allele frequencies. Six different previously defined *NAT1* star alleles, of varying frequencies, were called in the SSA populations (Table 1). *NAT1**4 (reference star allele—associated with slow acetylation) frequencies ranged from 27.3% (BFA) to 50.5% (LWK) (Table 1). The *4 global distribution was also common, with SSA having the lowest frequency of 42% and the EUR populations having the highest average frequency of 70%. A significant difference was found between these two biogeographical groups (*P* < 0.01) (Table S2). In addition, the *NAT1**4 SSA frequency differed significantly (*P* < 0.01) from the frequency in the SAS (58.7%) and AMR (61.2%) populations (Table 2, Table S2). The remaining *NAT1* slow acetylators star alleles (*3, *15, and *27) were either rare and/or absent in SSA populations (Table 2). Similarly, many of the *NAT1* slow acetylators star alleles (*14A, *14B, *17, *20, and *22) not called in SSA showed a low prevalence in the other global populations (Table 2).²⁹

*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 BRN (61.2%) and BFA (65.1%) participants (Table 1). In comparison with the other global populations included in this study, the *10 frequency in SSA (53.6%) was significantly higher (*P* < 3.5 × 10^{−3}) than in the African American/Afro-Caribbean (43.0%), AMR (35.3%), SAS (33.2%), and EUR (22.1%) populations (Table 2, Table S2). No significant difference in the frequencies of the *NAT1* star alleles nor any deviations from HWE were observed between the different SSA populations.

Twenty-four unique *NAT2* star alleles were called in the SSA populations analyzed in this study (Table 1). Two rapid *NAT2* acetylation star alleles, *NAT2**1 (reference haplotype) and *NAT2**4, were called across various global populations. The *1 frequency ranged from 9% (MAL) to 27.7% (BOT) (Table 1). The frequency in SSA (17.1%) was significantly higher compared with all the other global populations analyzed in this study (*P* < 0.001) (Table S2). In contrast, *4 occurred at 15.2% of the SSA population which was significantly lower (*P* < 1.2 × 10^{−4}) compared with AMR (22.6%) and EAS (51.0%) populations (Table 2, Table S2). *NAT2**5 was the most frequent slow acetylation star allele called in SSA (17.5%) and ranged from 5.1% (FNB) to 24.5% (BRN)

Table 2 Distribution of NAT1 and NAT2 star alleles in SSA compared with various global populations

Star alleles	Defining variants	Consequence	Acetylation effect	Allele frequencies (%) ^a					
				SSA (n = 1060)	African American/ Afro-Caribbean (n = 153)	AMR (n = 345)	EAS (n = 479)	EUR (n = 496)	SAS (n = 468)
NAT1									
*10	rs1057126 rs15561	3'-UTR 3'-UTR	Rapid	53.6	43.0	35.3	49.8	22.1	33.2
*11A	rs4986988	Intron	Rapid	0	0.3	1.0	0.3	1.7	3.0
	rs4986989	Intron							
	rs4987076	Missense (V149I)							
	rs4986990	Synonymous (T153T)							
	rs4986783	Missense (S214A)							
g.1077_1085del	3'-UTR	3'-UTR	2.8	0.3	0.3	0	1.4	2.0	4.4
	rs15561								
*3	rs15561	3'-UTR	Slow	42.0	50.2	61.2	45.9	70.0	58.7
*4	Reference	–	Slow	0	0.3	0.3	0	1.4	0
*14A	rs4986782 rs1057126 rs15561	Missense (R187Q) 3'-UTR 3'-UTR	Slow	0	0.3	0.4	0	0.9	0
*14B	rs4986782	Missense (R187Q)	Slow	0	0	0	0	0.6	0.3
15	rs5030839	Stop gain (R187)	No enzyme activity	0.1	0.3	0.4	0	0.9	0
*16	g.1091insAAA rs15561	Indel 3'-UTR	Unknown	0	0	0	1.9	0	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
*26B	g.1065_1090insTAA	Indel	Unknown	0.3	0	0	0	0	0
*27	rs4986992 rs4986991	Synonymous (L7L) Synonymous (S259S)	Slow	1.2	2.9	0.1	0	0	0
Allele frequencies (%) ^a									
Star alleles	Defining variants	Consequence(s)	Acetylation effect	SSA (n = 1076)	African American/ Caribbean (n = 157)	AMR (n = 346)	EAS (n = 504)	EUR (n = 503)	SAS (n = 488)
NAT2									
*1			Rapid	17.1	10.8	12.3	1.7	9.1	7.5
*4	rs1208	Missense (R268K)	Rapid	15.2	19.4	22.6	51.0	15.1	15.1
*5	rs1801280	Missense (I114T)	Slow	17.5	16.9	17.2	0.4	24.1	15.5
(Continued)									

(Continued)

Table 2 (Continued)

Star alleles	Defining variants	Consequence(s)	Acetylation effect	Allele frequencies (%) ^a					
				SSA (n = 1076)	African American/ Caribbean (n = 157)	AMR (n = 346)	EAS (n = 504)	EUR (n = 503)	SAS (n = 488)
*6	rs1799930 rs1208	Missense (R197Q) Missense (R268K)	Slow	12.9	14.6	11.3	23.9	18.7	24.9
*7	rs1208 rs1799931	Missense (R268K) Missense (G286E)	Slow	2.0	3.5	8.8	17.5	1.2	4.7
*10	rs72554617 rs1208	Missense (E167K) Missense (R268K)	Slow	0	0	0	0.1	0	0
*14	rs1801279 rs1208	Missense (R64Q) Missense (R268K)	Slow	10.1	6.7	0.3	0	0.1	0
*16	rs1801280 rs1208	Missense (I114T) Missense (R268K)	Slow	10.8	11.8	19.1	3.4	20.8	18.8
*30	rs1801280 rs1799930 rs1208	Missense (I114T) Missense (R197Q) Missense (R268K)	Slow	0.04	0	0	0	0	0
*40	rs1799931	Missense (G286E)	Slow	0.5	0.6	2.0	0.4	1.1	2.2
*49	rs1801280 rs79050330	Missense (I114T) Missense (T193M)	Slow	0.4	0.3	0	0	0	0.3
27	rs375746304 rs1208	Stop gain (R197) Missense (R268K)	Unknown	0.1	0	0	0	0	0
*31	rs45477599 rs1801280	Missense (L24I) Missense (I114T)	Unknown	0.4	0	0	0	0	0
*33	rs1801280 rs139351995	Missense (I114T) Missense (I158L)	Unknown	0.3	0.3	0	0	0	0
*34	rs1799930	Missense (R197Q)	Unknown	5.2	6.1	5.8	1.5	9.6	11.1
*36	rs12720065 rs1799930 rs1208	Missense (L135V) Missense (R197Q) Missense (R268K)	Unknown	1.5	0.6	0	0	0	0
*37	rs1799930 rs138707146 rs1208	Missense (R197Q) Missense (P213L) Missense (R268K)	Unknown	0.1	0	0	0	0	0
*38	rs1799930 rs55700793 rs1208	Missense (R197Q) Missense (K256E) Missense (R268K)	Unknown	1.0	0.6	0.1	0	0	0
*39	rs1799930 rs1208 rs56393504	Missense (R197Q) Missense (R268K) Missense (V280M)	Unknown	2.5	4.8	0	0	0	0.1
*41	rs79050330	Missense (T193M)	Unknown	0.2	0	0	0	0	0
*42	rs56387565	Missense (Y208H)	Unknown	0.04	0	0	0	0	0
*43	rs45618543	Missense (E203D)	Unknown	0.7	1.0	0.1	0	0	0
*44	rs12720065	Missense (L135V)	Unknown	1.0	0.6	0.1	0	0	0

(Continued)

Table 2 (Continued)

Star alleles	Defining variants	Consequence(s)	Acetylation effect	Allele frequencies (%) ^a				
				SSA (n = 1076)	African American/Afro-Caribbean (n = 157)	AMR (n = 346)	EAS (n = 504)	EUR (n = 503)
*45	rs72466456	Missense (I10T)	Unknown	0	0.3	0	0	0
*46	rs1801279	Missense (R64Q)	Unknown	0	0.3	0	0	0
*47	rs1801279 rs45518335 rs1208	Missense (R64Q) Missense (P228L) Missense (R268K)	Unknown	0	0.6	0	0	0
*50	rs200893121 rs1799930 rs1208	Missense (I3V) Missense (R197Q) Missense (R268K)	Unknown	0.04	0	0	0	0

AMR, admixed American participants; EAS, East Asian participants; EUR, European participants; SAS, South Asian participants; SSA, Sub-Saharan African participants; UTR, untranslated region.
^aAllele frequency estimates excluded individuals with unresolved diploypes. Given the novel alleles inferred in the study (see Table 6), the NAT1*4 and NAT2*1 frequencies should be considered as is and not “100 - (total frequencies of other alleles)”. The acetylation effects indicated are based on those reported in literature (ref 29) and the database of NATs (<http://nat.mbg.duth.gr/>). Further information (e.g. allele definitions) was obtained from PharmVar (<https://www.pharmvar.org/>; last accessed date: October 17, 2024) and PharmGKB (<https://www.pharmgkb.org/>; last accessed date: October 17, 2024).

(Table 1). The *5 frequency in SSA was significantly less frequent compared with the EUR populations (24.1%; $P < 0.001$) (Table S2). However, the frequency was significantly higher than the EAS populations (0.4%; $P < 0.001$).

Thirteen (57%) of the NAT2 star alleles called in SSA had an unknown functional activity (Table 1). Of these star alleles, only *34 was called in all the SSA populations analyzed in the study (1.8–12.1%), the remaining unknown function star alleles were absent in at least one of the SSA populations (Table 2). No significant frequency differences were observed for the NAT2 star alleles between the various SSA populations. However, when considering SSA populations jointly, deviations from HWE ($P < 0.02$) were noted for some NAT2 star alleles including *1, *4, *6, *14, and *39.

NAT1 and NAT2 phenotype predictions. The NAT1 intermediate acetylator phenotype was the most predominant predicted phenotype in SSA (50.6%), ranging from 42.9% (MAL) to 56.2% (YRI) (Figure 1a). However, in general, NAT1 phenotype distributions were non-uniform across the SSA populations in this study. For example, the NAT1 slow acetylation phenotype (mainly accounted for by *4/*4) was significantly less frequent in SSA (20.8%), especially in comparison with AMR (42.9%), SAS (38.7%), and EUR (58.8%) populations (Figure 1b). The MAL and CAM populations had the highest frequency (34.7%) of the NAT1 rapid acetylator phenotype (Figure 1a). The frequency in SSA was 28.4% and was significantly ($P < 2.4 \times 10^{-3}$) higher in most of the global populations, apart from the EAS populations (Table S3). The only diplotype associated with this phenotype in the various global populations in this study was *10/*10 (Table S4).

Similar to the SSA NAT1 acetylator phenotypes, the NAT2 intermediate acetylator was most common in SSA (52.7%). The BOT population had the highest NAT2 intermediate acetylator frequency (67.5%) across the SSA populations analyzed in this study (Figure 1a). The frequency of this phenotype was significantly higher in SSA when compared with the EUR populations ($P < 2.3 \times 10^{-2}$) (Table S3). NAT2 diplotypes which accounted for this phenotype were observed at relatively low frequencies individually in SSA (Table S5). Between 23.3% and 49.9% of individuals across the global populations were predicted to be NAT2 slow acetylators (Figure 1b). In the SSA populations, 35.7% of the participants were predicted to be NAT2 slow acetylators, significantly lower ($P < 0.01$) compared with the 48.4% in EUR and the 49.9% in SAS populations (Table S3). The *5/*5 diplotype contributed most to the NAT2 slow acetylator phenotype in each of the populations, except in the EAS and SAS where *6/*7 and *6/*6 were the predominant diplotypes, respectively (Table S5). The NAT2 rapid acetylator phenotype was only observed in 11.5% of the SSA participants. In most of the global populations, the NAT2 rapid acetylator was less frequent, with EUR (8.4%) and SAS (7.7%) populations having lower frequencies than what was observed in SSA for this phenotype (Figure 1). The only exception was in the EAS populations, where 32% of individuals were NAT2 rapid acetylators. This was significantly higher when compared with SSA ($P < 0.01$) (Table S3). The phenotype was mainly attributable to the *4/*4 diplotype across SSA and the other global populations (Table S5).

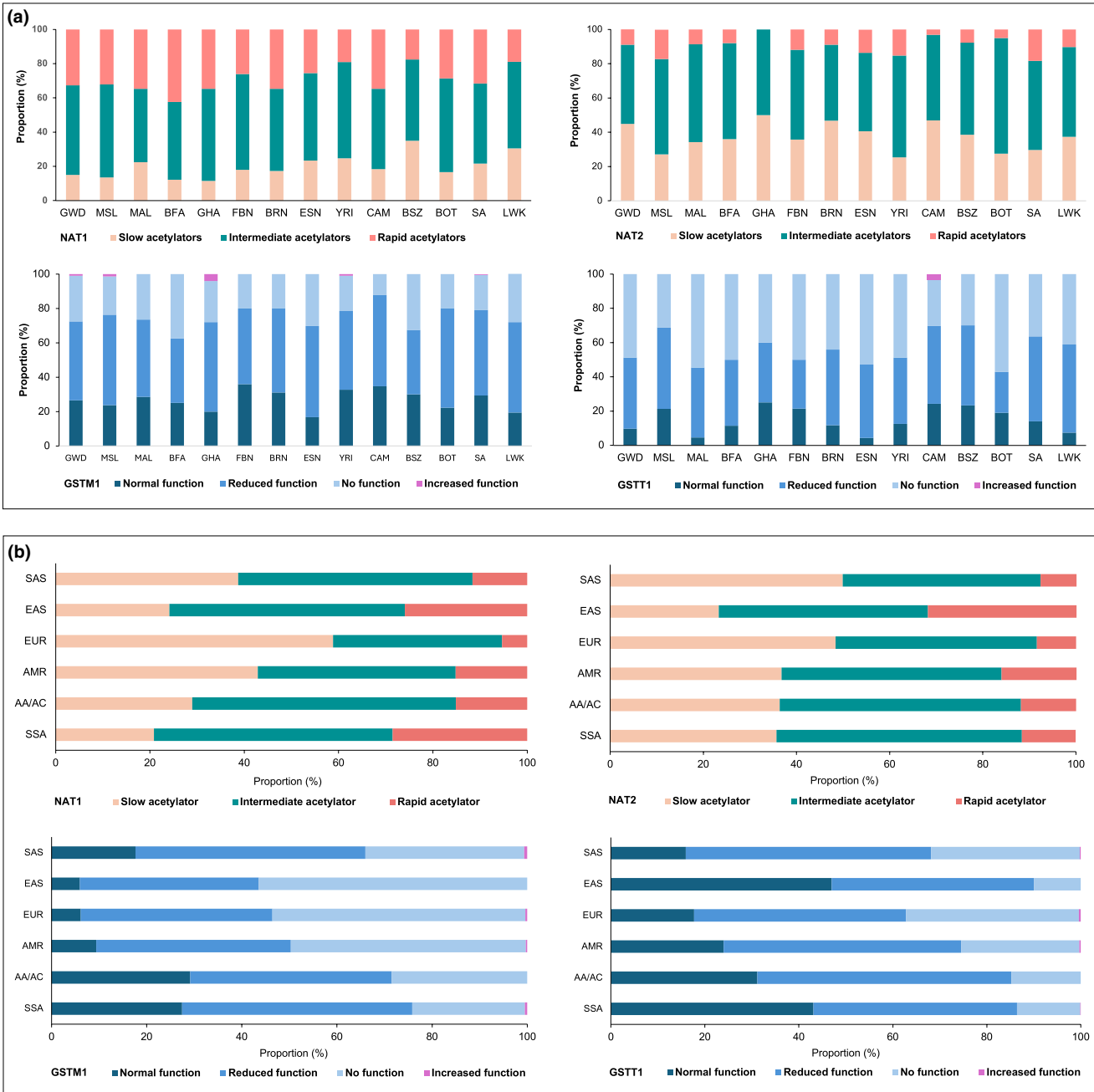


Figure 1 Distribution of NAT1, NAT2, GSTM1, and GSTT1 phenotype predictions (based on diplotypes). (a) Predicted phenotype distributions in 14 SSA populations and (b) in comparison with other global populations. AA/AC, African American/Afro-Caribbean participants; AMR, admixed American participants; BFA, Burkina Faso participants; BOT, Botswana participants; BRN, Berom of Nigeria; BSZ, Bantu speakers from Zambia; CAM, Cameroonian participants; EAS, East Asian participants; ESN, Esan in Nigeria; EUR, European participants; FNB, Fon from Benin; GHA, Ghanaian participants; GWD, Gambian in Western Division Mandinka; LWK, Luhya in Webuye, Kenya; MAL, Malian participants; MSL, Mende in Sierra Leone; SA –South African participants; SAS, South Asian participants; YRI, Yoruba in Ibadan (Nigeria).

GSTM1 and GSTT1 star allele frequencies. The *GSTM1**0 allele (gene deletion) was common across the various SSA populations ranging from 42% in FNB to 57.4% in the ESN population (Table 3). A higher prevalence of *GSTM1**0 was observed among the global populations, especially in EUR (73.5%) and EAS (75.3%) populations (Table 4). Both these frequencies were significantly higher ($P < 0.001$) compared with the average frequency in SSA (48.3%) (Table S2). *GSTM1**A (associated with normal GSTM1

enzyme activity) was relatively frequent in the SSA populations (38.5–52.5%) (Table 3). When comparing to the other global populations, the *A distribution was similar between SSA and the African American/Afro-Caribbean populations (Table 4). When compared to the remaining global populations, SSA had a significantly higher *A frequency. Both *A_{x2} and *B_{x2} were relatively rare, with *B_{x2} only being called in the SAS populations (Table 4).

Table 3 Distribution of *GSTM1*, *GSTT1* and *GSTP1* variation across SSA populations in this study

Star alleles	Detoxification activity	Allele frequency (%) ^a														
		All (n=1061)	GWD (n=111)	MSL (n=81)	MAL (n=50)	BFA (n=33)	GHA (n=25)	FNB (n=50)	BRN (n=47)	ESN (n=98)	YRI (n=108)	CAM (n=50)	BSZ (n=41)	BOT (n=46)	SA (n=198)	LWK (n=96)
GSTM1																
*0	None	48.3	50.0	48.4	49.5	56.9	49.0	42.0	44.7	57.4	43.7	38.4	51.9	50.0	45.3	55.2
*A	Normal	47.2	45.9	48.4	46.5	38.5	47.1	49.0	52.1	39.5	51.2	52.5	46.9	47.8	47.9	42.7
*B	Normal	4.3	3.6	2.5	4.0	4.6	2.0	9.0	3.2	3.1	4.7	9.1	1.2	2.2	6.5	2.1
*Ax2	Increased	0.2	0.5	0	0	0	2.0	0	0	0	0.5	0	0	0	0.3	0
Star alleles	Detoxification activity	Allele frequency (%) ^a														
		All (n=890)	GWD (n=96)	MSL (n=73)	MAL (n=47)	BFA (n=29)	GHA (n=23)	FNB (n=46)	BRN (n=41)	ESN (n=81)	YRI (n=92)	CAM (n=40)	BSZ (n=36)	BOT (n=32)	SA (n=145)	LWK (n=81)
GSTT1																
*0	None	65.0	60.9	57.5	75.5	70.0	72.3	62.0	65.4	73.5	67.8	50.0	39.0	67.2	62.3	63.4
*A	Normal	35.0	39.1	42.5	23.4	31.0	27.7	38.0	34.6	26.5	32.2	46.3	32.0	32.8	37.7	36.6
*Ax2	Increased	0.2	0	0	1.1	0	0	0	0	0	0	3.8	0	0	0	0
Variant		Allele frequencies (%) ^a														
		All (n=1079)	GWD (n=113)	MSL (n=85)	MAL (n=50)	BFA (n=33)	GHA (n=26)	FNB (n=50)	BRN (n=49)	ESN (n=99)	YRI (n=108)	CAM (n=50)	BSZ (n=41)	BOT (n=47)	SA (n=199)	LWK (n=99)
GSTP1																
rs1695 (I105V) ^b		46.7	53.5	47.1	61.0	42.4	40.4	50.0	34.7	54.0	39.8	40.0	39	46.8	51.3	51.0
rs533546581 (S106F)		0.05	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0
rs1138272 (A114G) ^b		0.2	0	0	0	0	0	0	0	0	0	0	0	1.1	0.3	1.5
rs553211668 (K141R)		0.05	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0
rs199949339 (I144T)		0.2	0	0	0	0	3.8	0	0	1.0	0	0	0	0	0	0
rs201449696 (I162L)		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0
rs188653023 (R183H)		0.05	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0
rs45549733 (R187W)		0.1	0	0	0	0	0	2.0	0	0	0	0	0	0	0	0
rs780366806 (Glu198GlyfsTer34)		0.05	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0
rs376074280 (A201T)		0.05	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0

BFA, Burkina Faso participants; BOT, Botswana participants; BRN, Berom of Nigeria; BSZ, Bantu speakers from Zambia; CAM, Cameroonian participants; ESN, Esaan in Nigeria; FNB, Fon from Benin; GHA, Ghanaian participants; GWD, Gambian in Western Division Mandinka; LWK, Luhya in Webuye, Kenya; MAL, Malian participants; MSL, Mende in Sierra Leone; SA, South African participants; YRI, Yoruba in Ibadan (Nigeria). ^aAllele frequencies calculated with excluding novel haplotypes and individuals with unresolved diplotypes. ^bSNVs used to define known *GSTP1* star alleles including *B (rs1695), *C (rs1695, rs1138272), and *D (rs1138272).

Table 4 Distribution of *GSTM1*, *GSTT1* and *GSTP1* variation in SSA compared with various global populations

				Allele frequency (%) ^a					
Star alleles	Defining variants	Consequence	Detoxification activity	SSA (n = 1061)	African American/ Afro-Caribbean (n = 154)	AMR (n = 343)	EAS (n = 504)	EUR (n = 495)	SAS (n = 483)
GSTM1									
*O	–	Gene deletion	None	48.3	50.5	69.8	75.3	73.5	57.7
*A	Reference	–	Normal	47.2	45.0	18.7	9.1	17.7	32.9
*B	rs1065411	Missense (K173N)	Normal	4.3	4.6	11.4	15.6	8.6	9.0
*Ax2		Gene duplication	Increased	0.2	0	0.1	0	0.2	0.3
*Bx2		Gene duplication	Unknown	0	0	0	0	0	0.1
				Allele frequency (%) ^a					
Star alleles	Defining variants	Consequence	Detoxification activity	SSA (n = 890)	African American/ Afro-Caribbean (n = 139)	AMR (n = 343)	EAS (n = 498)	EUR (n = 500)	SAS (n = 483)
GSTT1									
*O	–	Gene deletion	None	65.0	58.1	49.1	68.7	40.2	42.1
*A	Reference	–	Normal	35.0	41.9	50.8	31.3	59.6	57.8
*Ax2		Gene duplication	Unknown	0.2	0	0.1	0	0.2	0.1
			Allele frequencies (%)						
rsID	Consequence	SSA (n = 1079)	African American/ Afro-Caribbean (n = 157)	AMR (n = 347)	EAS (n = 504)	EUR (n = 503)	SAS (n = 489)		
GSTP1									
rs560001291	Missense (R14C)	0	0	0.1	0	0	0	0	
rs568270040	Frameshift (Ala23ArgfsTer80)	0	0	0	0	0	0	0.3	
rs551651791	Stop gain (W29*)	0	0	0	0	0	0	0.1	
rs192307201	Missense (A23S)	0	0	0.6	0	0.2	0	0	
rs753269233	In-frame deletion (Glu41del)	0	0	0	0.2	0	0	0	
rs202114011	Missense (M92T)	0	0	0	0	0.1	0	0	
rs758425456	Missense (G96S)	0	0	0.1	0	0	0	0	
rs1695 ^b	Missense (I105V)	46.7	44.6	47.3	17.9	33.0	29.5		
rs533546581	Missense (S106F)	0.1	0	0	0	0	0	0	
rs563603340	Missense (Y112C)	0	0	0	0.4	0	0	0	
rs1138272 ^b	Missense (A114G)	0.2	1.3	2.9	0.1	7.1	7.1		
rs78507509	Missense (P124A)	0	0	0	0.1	0	0	0	
rs71534227	Missense (G139E)	0	0	0	0	0.1	0	0	
rs553211668	Missense (K141R)	0.1	0	0	0	0	0	0	
rs199949339	Missense (I144T)	0.2	0.6	0	0	0	0	0	
rs4986949	Missense (D147Y)	0	0	0	0.1	0	0	0	
rs199567349	Missense (L160M)	0	0	0.1	0	0	0	0	
rs201449696	Missense (I162L)	0.1	0.3	0	0	0	0	0	
rs188653023	Missense (R183H)	0.1	0	0.3	0	0	0	0	
rs45549733	Missense (R187W)	0.1	0	0	0	0	0	0	
rs191595383	Missense (P197A)	0	0	0	0	0.1	0	0	
rs780366806	Frameshift (Glu198GlyfsTer34)	0.1	0	0	0	0	0	0	
rs376074280	Missense (A201T)	0.1	0	0	0	0	0	0	

AMR, admixed American participants; EAS, East Asian participants; EUR, European participants; SAS, South Asian participants; SSA, Sub-Saharan African participants. ^aAllele frequencies calculated excluding individuals with unresolved diplotypes. ^bSNVs used to define known *GSTP1* star alleles including *B (rs1695), *C (rs1695, rs1138272), and *D (rs1138272).

Both *GSTT1*0* and *GSTT1*A* occurred at high frequencies of 65% and 35%, respectively, in SSA (Table 3). *GSTT1*0* was most prevalent, especially in GHA (72.3%), ESN (73.5%), and MAL (75.5%) (Table 3). There were no significant differences in the **0* and **A* frequencies among the SSA populations. However, a deviation from HWE was observed in the FNB population ($P=0.02$). The high *GSTT1*0* frequency was observed in both the SSA and the other comparative global populations (Table 4). However, the **0* frequency in SSA differed significantly from the AMR (49.1%), EUR (40.2%), and SAS (42.1%) populations (Table S2). The *GSTT1*A* frequency in SSA (35%) was significantly lower ($P < 9.3 \times 10^{-12}$) when compared to the frequencies among the AMR (50.8%), EUR (59.6%), and SAS (57.8%) populations (Table S2). The **A*2 duplication was only called in the SSA, AMR, EUR, and SAS populations at low frequencies (0.1–0.2%) (Table 4).

GSTM1 and GSTT1 phenotype predictions. In the SSA populations, 27.5% of participants had a predicted normal detoxification activity (two normal function alleles). 24.4% of participants had a homozygous *GSTM1* deletion (results in no production of the enzyme). The *GSTM1* no function phenotype was highest in BSZ (32.5%) and BFA (37.5%) (Figure 1a). This frequency was significantly ($P < 0.01$) different from that observed in the EUR (53%) and EAS (56%) populations (Table S3). The reduced *GSTM1* activity phenotype was predominant among the various global populations, especially in SSA (49%) and SAS (48%) (Figure 1b). This phenotype ranged from 37.5% in BSZ and BFA to 57.8% in BOT. A significant difference between SSA and EAS (38%) was observed ($P < 0.01$) (Table S3). 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 contributed more to the reduced *GSTM1* activity in the EUR populations (Table S6).

For *GSTT1*, 13.4% of SSA participants had a predicted normal detoxification activity (two normal function alleles). Predicted absence of *GSTT1* activity (inferred from the **0/*0* diplotype) ranged from 27.3% in CAM to 57.1% in BOT (Figure 1a). This phenotype was largely prevalent in SSA (43.1%) as well as in the other global populations such as African American/Afro-Caribbean (31.1%), and EAS (47%) (Figure 1b). The frequency of this phenotype was significantly higher ($P < 0.01$) in SSA (43.1%) compared with the frequencies observed in the AMR (24%), EUR (18%) and SAS (16%) populations (Table S3). The reduced *GSTT1* activity phenotype was the most prevalent *GSTT1* phenotype in the SSA populations (54.1%) and ranged from 23.8% in BOT to 49.5% in SA (Figure 1). The high prevalence of this phenotype was observed in other global populations such as in the AMR, SAS and EUR populations (Figure 1b); however, there were no significant differences in the frequency between the various global populations (Table S3). The SSA, African American/Afro-Caribbean, and EAS populations had lower frequencies of the normal *GSTT1* phenotype, compared with the remaining three populations. The distribution of this phenotype was significantly lower ($P < 4.1 \times 10^{-5}$) in SSA (13%) compared with the AMR (25%), SAS (32%), and EUR (37%) populations (Table S7).

The null deletion of both *GSTM1* and *GSTT1* was observed in 6.8% of the SSA participants. The deletion was observed at a slightly higher frequency compared with the presence of both genes (5.5%) (Table S8).

GSTP1 variation across African populations. Ten nonsynonymous variants were called across the SSA populations, this included nine missense variants and one frameshift variant (Table 3). The rs1695 was most frequent in the SSA populations ranging from 34.7% (BRN) to 61.0% (MAL) (Table 4). In comparison with other global populations, the frequency in SSA (46.7%), was significantly higher ($P < 0.001$) than the frequency observed in the EAS (17.9%), SAS (29.5%), and EUR (33.0%) populations (Table 3). Three *GSTP1* variants were predicted to be deleterious by the *in silico* tools used in this study (Table S13); of these, rs45549733 and rs780366806 were called only in the SSA populations (Table 3).

At present, there are four *GSTP1* star alleles: **A* (reference), **B* (rs1695), **C* (rs1695, rs1138272), and **D* (rs1138272). The *GSTP1*B* frequency was most frequent in SSA (95.8%) and differed significantly ($P < 0.001$) from the AMR (88.2%), EUR (55.1%), and SAS (46.8%) populations (Tables S2 and S9). Nine *GSTP1* haplotypes, not previously used to define *GSTP1* star alleles, were estimated in the SSA populations (Table S9). The most frequent was rs1695~A>G; rs201449696~A>C (0.5%), which was also observed in the Africa American/Afro-Caribbean populations (0.6%) (Table S9).

CYP2E1 variation across African populations. Unlike other pharmacogenomic-relevant *CYP* genes, *CYP2E1* has not yet been included in PharmVar. A few star alleles (**1A* to **7C*) have been published, 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* nonsynonymous variants, possible novel haplotypes, and structural variants, without particular focus on the star allele nomenclature.

Missense and stop-gain variants were the only two types of *CYP2E1* nonsynonymous variants identified in the SSA and other global populations (Table 5, Table S10). The g.5167T>C (p.Leu45Pro) variant was identified only in the SA population (0.3%) and was absent in the remaining SSA and other global population groups (Table 5, Table S10). It was predicted to be deleterious by the majority of the *in silico* predictions tools used in this study (Table S13). The absence of this missense variant on dbSNP and gnomAD indicates that it may be a novel variant.

Of the remaining 11 missense variants identified in the various SSA populations, rs6413419 (defines **4*) was relatively common among the SSA populations ranging from 13.6% (LWK) to 55.0% (SA) (Table 5). The frequency in SA was statistically different from that observed in the LWK population. The SSA frequency (24.3%) and was statistically different ($P < 0.001$) from the AMR, EUR, and SAS populations (Tables S2 and S10).

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 > 16% in the SSA

Table 5 Distribution of CYP2E1 variants across SSA populations in this study

rsID	Consequence	SSA (n = 1079)	Allele frequencies (%)										BOT (n = 47)	BSZ (n = 41)	SA (n = 198)	LWK (n = 99)
			GWD (n = 113)	MSL (n = 85)	MAL (n = 50)	BFA (n = 33)	GHA (n = 26)	FNB (n = 50)	BRN (n = 49)	ESN (n = 99)	YRI (n = 99)	CAM (n = 50)				
Nonsynonymous variants																
rs367957731	Missense (L17F)	0.1	0.4	0	1	0	0	0	0	0	0	0	0	0	0	0
g.5167T>C	Missense (L45P)	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0
^a rs6413419	Missense (V179I)	24.3	21.7	24.1	17.0	31.8	19.2	26.0	28.6	28.8	31.9	25.0	23.2	20.2	55.0	13.6
rs555344732	Stop gain (Y245*)	0.1	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0
rs749979006	Missense (R282H)	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0
rs1007026512	Missense (D295N)	0.1	0	0	0	1.5	0	0	0	0	0	0	0	0	0	0
rs759707874	Missense (R337Q)	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rs372279263	Missense (T362I)	0.1	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0
rs59981143	Missense (T373A)	0.1	0.4	0	0	0	0	0	0	0	0.5	0	0	0	0	0
rs59656378	Missense (V396L)	0.4	0.4	0	0	0	0	0	1.0	0	0.5	1.0	3.7	1.1	0.3	0
rs57702102	Missense (N418S)	0.2	0	0	0	3.0	0	2.0	0	0	0	0	0	0	0	0
rs142767992	Missense (S424G)	0.1	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0
rs199855848	Missense (M445I)	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
rs28969387	Missense (H457L)	5.2	6.2	5.9	11.0	9.1	5.8	5.0	6.1	3.5	6.0	6.0	4.9	5.3	4.0	2.5
Non-coding variants defining in previous CYP2E1 star allele definitions																
^a rs3813867	2 KB Upstream	5.2	8.0	6.5	4.0	13.6	6.5	5.0	2.0	8.6	8.3	4.0	1.2	0	3.5	2.5
^a rs2070672	2 KB Upstream	16.8	21.7	18.2	12.0	12.1	17.7	12.0	17.3	10.6	16.2	17.0	14.6	29.8	17.8	10.6
^a rs2070673	2 KB Upstream	21.1	24.3	21.2	16.0	18.1	24.2	22.0	14.3	23.2	19.4	21.0	17.1	12.8	24.8	19.2
^a rs6413420	2 KB Upstream	1.2	3.5	1.2	1.0	0	0	0	2.0	1.5	0.5	2.0	4.9	0	0.3	1.0
^a rs2070676	Intron	26.4	32.3	25.9	25.0	33.3	30.6	21.0	21.4	31.8	33.8	30.0	11.0	21.3	22.0	23.7

BFA, Burkina Faso participants; BOT, Botswana participants; BRN, Berom of Nigeria; BSZ, Bantu speakers from Zambia; CAM, Cameroonian participants; ESN, Esaan in Nigeria; FNB, Fon from Benin; GHA, Ghanaian participants; GWD, Gambian in Western Division Mandinka; LWK, Luhya in Webuye, Kenya; MAL, Malian participants; MSL, Mende in Sierra Leone; SA, South African participants; YRI, Yoruba in Ibadan (Nigeria). ^aSNVs used to define known CYP2E1 star alleles including *1B (rs2070676 (30)), *2 (rs72559710), *3 (rs55897648) (31), *4 (rs6413419) (32), *5A (rs3813867, rs2031920) (33,34), *5B (rs3813867, rs2031920) (32), *7A (rs2070676), *7B (rs6413420, rs2070676), and *7C (rs2070672, rs2070676) (32)—definitions may be incomplete due to inadequate gene sequencing in previous studies. The g.13007T>A variant defines *5A and *6A (34), and it was not called in this study.

populations (Table 5). Significantly higher frequencies for both variants were observed across the global populations, particularly in the EUR populations where the rs2070673 and rs2070676 variants occurred at 81.6% and 87.2%, respectively (Table S10). 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 S10).

Eighteen *CYP2E1* haplotypes, defined by nonsynonymous variants, were identified in the SSA populations; the highest number called in comparison with the other global populations (Table S11). The reference haplotype (*CYP2E1* sequence in GRCh38) was observed at a frequency of 41.3% in SSA and 45.2–95.8% in the other global populations. The frequency in SSA was statistically lower in comparison with other population groups, except for the African American/Afro-Caribbean populations. The haplotype defined by rs6413419~G>A was significantly higher ($P < 0.001$) in SSA (46.5%) compared with the AMR (8.9%), EUR (6.2%), and SAS (2.2%) populations (Table S2).

We identified *CYP2E1* duplications (copy number > 2) at a frequency of 5.64% in SSA; this was higher in comparison with other global populations such as the EUR populations (2.98%) (Figure S4). The deletion of *CYP2E1* was observed at a frequency of 0.5% in SSA. The null deletion of *CYP2E1* was not observed in SSA; however, it was observed in the AMR populations.^{30–34}

Novel haplotypes

Forty-seven possible resolved novel haplotypes were identified across *NAT1*, *NAT2*, *GSTM1*, and *GSTT1* (Table 6); 26% of which haplotype 1, 9, 10, 24, 25, 26, 28, 32, 39, 41, 43, and 44) were defined by variants predicted to be deleterious, including rs554793519, rs529845573, rs138061602, rs561124342, rs375154572, rs184653774, and rs533860247, using a consensus of SIFT, CADD, Polyphen-2, and PROVEAN (Table S13). These variants were prioritized for further follow-up with the structure-based predictor DynaMut2. From this analysis, *NAT1* rs554793519, rs529845573, rs138061602, *NAT2* rs561124342, and *GSTM1* rs375154572, rs184653774, and rs533860247 were predicted to destabilize ($\Delta\Delta G < 0$ kcal/mol) the respective proteins. Novel *NAT1* and *NAT2* haplotypes identified in the 1000 Genomes Project dataset were further resolved using trio data, where available (Table S12). Of these, *NAT2**53 (Figure 2) has been reviewed by the PharmVar *NAT2* expert panel and is now cataloged in the PharmVar database (<https://www.pharmvar.org/gene/NAT2>).

DISCUSSION

The high TB burden in Africa emphasizes a need for effective preventative measures and treatments. However, with the considerable occurrence of ADRs to anti-TB drugs, in particular AT-DILI,^{6,35} and treatment resistance, it is necessary to optimise drug dosage/choice to improve drug safety and efficacy. One major obstacle in adopting this precision medicine approach in Africa is the lack of sufficient pharmacogenomic profiling of function-altering star alleles influencing drug response. We analyzed 1079 high-coverage genomes of participants from various SSA ethnolinguistic groups, to establish a more comprehensive insight

into the pharmacogenomic diversity in six important pharmacogenes (*NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *CYP2E1*, and *GSTP1*), which are relevant for response to medicines for TB, cancer, and/or hypertension, among other diseases 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.

As evidenced by the number of variants per kilobase, *NAT1*, *NAT2*, *GSTM1*, *GSTT1*, *CYP2E1*, and *GSTP1* showed higher genetic diversity across SSA populations compared with the other global populations included in this study. Of the nonsynonymous variants identified in these genes, > 30% were predicted to be deleterious which warrants their further *in vitro* and/or *in vivo* functional investigation to support their potential inclusion in pharmacogenetic testing panels. Additionally, we identified a high number of potentially novel haplotypes ($n = 47$) across four pharmacogenes (*NAT1*, *NAT2*, *GSTM1*, and *GSTT1*). This can be attributed to the fact that we analyzed high-coverage genomes from various underrepresented SSA populations.

Based on the diplotypes called in this study, SSA (and EAS) populations had a higher proportion of individuals with a rapid acetylator phenotype (mostly due to the *10 distribution) compared with other biogeographical groups while the slow acetylator phenotype was more frequent in EUR populations. This is consistent with previous observations and could be accounted for by variability in diet and environmental selection pressures.^{36,37}

There is scarcity of research characterizing the influence *NAT1* star alleles may have on the response to anti-TB drugs. However, *NAT1**3 and *14 have been associated with a higher exposure to p-Aminosalicylic acid.³⁸ For other treatments, *NAT1**10 has been associated with decreased concentration of thioguanine nucleotides consequently leading to increased toxicity and decreased efficacy of thiopurines in individuals with inflammatory bowel disease.³⁹

Regarding *NAT2* variation, the SSA populations had a higher number of unique *NAT2* star alleles called in comparison with the other global populations in this study. In general, *NAT2* slow acetylator star alleles, in particular *5 (rs1801280) and *6 (rs1799930 and rs1208), were more prevalent in both SSA and other global populations, compared with rapid acetylator star alleles. This has previously been attributed to positive selection as population groups transitioned from being hunter gathers to pastoralists and farmers.^{25,40} There have been several studies reporting associations between *NAT2* variation, particularly the slow acetylator profile, with anti-TB drug-induced hepatotoxicity or liver injury (DILI) via increased exposure to INH and hydrazine.^{41,42} These ADRs are either due to INH (a hydrazine derivative) itself and/or its metabolites, including acetyl-hydrazine, hydrazine and ammonia, presumably from the formation of reactive oxygen species that can cause necrosis and autoimmunity.⁴¹ However, there are some reports that have found no significant association.⁴³ That being said, there is a paucity of data from high TB burden countries in SSA and beyond, which could elucidate population specific pharmacogenetic, pharmacokinetic, and pharmacodynamic profiles. Future research under the SAMRC/GSK/Norvatis-supported GRADIENT initiative⁴⁴ and other projects could yield critical data to enhance our understanding of the impact of African genetic heterogeneity in *NAT2*

Table 6 Possible novel haplotypes identified across SSA populations in this study

#	Background allele	Additional core variant	Consequence	Allele count	Population
NAT1					
1	*4	rs554793519~g.57068T>C	Missense (L161P)	1	Gambia (1000G)
2	*4	rs201916736~g.57277G>T	Missense (V231F)	1	Nigeria (1000G)
3	*4	g.57013~C>T	Stop gain (Q143*)	1	South Africa (CBRL)
4	*4	rs4986992~g.56607T>G	Synonymous (L7L)	4	Nigeria (1000G, SGDP) Cameroon (H3Africa)
5	*4	rs112846752	Synonymous (L252L)	1	Botswana (H3Africa)
6	*4	g.57669delAAT	3'-UTR	3	South Africa (AWI-Gen, SAHGP)
7	*3	g.57669ins(AAT) ₃	3'-UTR	1	Gambia (1000G)
8	*10	rs111848753~g.56684G>A	Missense (R33Q)	3	South Africa (AWI-Gen, SAHGP, CBRL)
9	*10	rs138061602~g.56777G>A	Missense (R64Q)	1	Nigeria (1000G)
10	*10	rs529845573~g.57025C>T	Missense (P147S)	1	Gambia (1000G)
11	*10	rs1389255907~g.57437G>A	Missense (G284D)	1	South Africa (AWI-Gen)
12	*10	rs4986991~g.57363T>C	Synonymous (S259S)	1	Kenya (1000G)
13	*10	rs15561~g.57681A>C	3'-UTR	3	South Africa (CBRL, SGDP)
14	*10	g.57666delAATAATAAA	3'-UTR	3	Kenya (1000G)
15	*10	g.57669insAAT	3'-UTR	1	Nigeria (1000G)
16	*10	g.57669insAATAATAAT	3'-UTR	1	Sierra Leone (1000G)
17	*10	rs548004925~56972A>G+ rs15561~57681A>C	Missense (Y129C) 3'-UTR	1	Sierra Leone (1000G)
18	*10	rs932721238~g.57671delT(A)5	3'-UTR	1	Sierra Leone (1000G)
19	*10	rs370302501-18222499~G>A	Missense (R151H)	1	Sierra Leone (1000G)
20	*10	g.57669insAAT	3'-UTR	1	Nigeria (1000G)
21	*10	rs15561~57681A>C	3'-UTR	1	Nigeria (1000G)
22	*10	g.57665del(TAA)3 rs15561~57681A>C	3'-UTR	2	Mali, Zambia (H3Africa)
23	*10	rs370859724~g.56777G>A	Missense (M105I)	1	South Africa (SGDP)
NAT2					
24	*6	rs561124342~g.14007G>C	Missense (G83A)	1	Nigeria (1000G)
25	*14	rs149460636~g.14466G>A	Missense (G236D)	1	Benin (H3Africa)
26	*14	rs374761885~g.14328A>G	Missense (T190C)	1	Cameroon (H3Africa)
GSTM1					
27	*A	rs449856~g.10471T>A	Missense (S210T)	15	South Africa (AWI-Gen, CBRL) Mali, Botswana, Zambia (H3Africa) Nigeria, Kenya, Sierra Leone (1000G)
28	*A	rs375154572~g.7704C>T	Missense (R168C)	6	Sierra Leone (1000G)
29	*A	rs184653774~g.5105C>A	Missense (D9E)	4	Nigeria (H3Africa) Kenya (1000G) South Sudan (SGDP)
30	*A	rs146668816~g.6330C>G	Missense (N85K)	3	Gambia, Kenya (1000G)
31	*A	rs138440339~g.7677C>G+ rs1065411~g.7721G>C	Missense (L159F) Missense (K173N)	3	Gambia, Sierra Leone (1000G)
32	*A	rs533860247~g.10480G>A	Missense (A213T)	3	Nigeria (1000G, H3Africa)
33	*A	rs138440339~g.7677C>G	Missense (L159V)	1	Cameroon (H3Africa)
34	*A	rs146668816~g.6330~C>G+ rs1065411~g.7721~G>C	Missense (N85K) Missense (K173N)	1	Sierra Leone (1000G)

(Continued)

Table 6 (Continued)

#	Background allele	Additional core variant	Consequence	Allele count	Population
35	*A	rs138440339~g.7677~C>G+ rs1065411~g.7721~G>C (CN=4)	Missense (L159F) Missense (K173N)	1	Ghana (H3Africa)
36	*B	rs449856~g.10471T>A	Missense (S210T)	1	Burkino Faso (H3Africa)
<i>GSTT1</i>					
37	*A	rs2266637~g.271020C>T+ rs8140585~g.277188C>T	Missense (V169I) Missense (A61T)	209	Benin, Botswana, Cameroon, Ghana, Mali, Zambia, Nigeria (H3Africa) Nigeria, Gambia, Kenya (1000G) South Africa (AWI-Gen, SAHGP, CBRL) Nigeria (SGDP)
38	*A	rs8140585~g.277188C>T	Missense (A61T)	73	Benin, Botswana, Cameroon, Ghana, Mali, Zambia, Nigeria (H3Africa) Nigeria, Gambia, Kenya, Sierra Leone (1000G) South Africa (AWI-Gen, CBRL) Gambia (SGDP)
39	*A	rs2266637~g.271020C>T	Missense (V169I)	65	Benin, Botswana, Cameroon, Ghana, Mali, Zambia, Nigeria (H3Africa) Nigeria, Gambia, Kenya, Sierra Leone (1000G) South Africa (AWI-Gen, SAHGP, CBRL) South Africa, Kenya (SGDP)
40	*A	rs1601976046~g.270731C>T+ rs8140585~277188C>T	Missense (R197H) Missense (A61T)	12	Botswana (H3Africa) South Africa (AWI-Gen, SAHGP, CBRL)
41	*A	rs141759372~g.273586A>G	Missense (W101R)	9	Cameroon (H3Africa) Nigeria, Sierra Leone (1000G) South Africa (AWI-Gen)
42	*A	rs1601976046~g.270731C>T	Missense (R197H)	5	Botswana (H3Africa) South Africa (AWI-Gen, CBRL)
43	*A	rs144686326~g.270759-G>A	Stop gain (R188*)	3	Burkino Faso, Nigeria (H3Africa) Sierra Leone (1000G)
44	*A	rs144686326~g.270759-G>A rs141759372~g.273586A>G	Stop gain (R188*) Missense (W101R)	2	Nigeria (1000G)
45	*A	rs199521920~g.275948C>T	Missense (D43N)	2	Botswana (H3Africa) Kenya (1000G)
46	*A	rs1601976046~g.270731C>T+ rs2266637~g.271020C>T+ rs8140585~g.277188C>T	Missense (R197H) Missense (V169I) Missense (A61T)	2	South Africa (CBRL, SAHGP)
47	*A	rs2266637~g.271020C>T+ rs8140585~g.277188C>T+ g.271162_271163del	Missense (V169I) Missense (A61T) Frameshift (Pro121fs)	1	Nigeria (1000G)

AWI-Gen, Africa Wits-INDEPTH partnership for Genomics; CBRL (NICD, WITS), Cell Biology Research Lab, National Institute of Communicable Diseases, University of the Witwatersrand; H3Africa, Human Heredity & Health in Africa; SAHGP, South African Human Genome Project; SGDP, Simons Genome Diversity Project.

(and other genes) on the response to INH and other key anti-TB drugs. This is highly relevant in Africa where we observed an estimated 25–50% of participants with a NAT2 slow acetylator profile.

The high prevalence of *GSTM1**0 (42.0–57.4%) and *GSTT1**0 (39.0–75.5%) in the various SSA study populations is remarkable given the importance of *GSTM1* and *GSTT1* in detoxification of metabolites and toxins in the liver.⁴⁵ The high occurrence of these two gene deletions is comparable to other SSA-based population

studies^{46,47}; however, there are differences in their distribution as noted in this study. 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, which is supported by results in previous studies.^{46,48,49} In keeping with these observations, the *GSTT1* null phenotype is more frequent across SSA populations compared with EUR populations while the *GSTM1* null phenotype is more frequent in EUR populations. The distributions

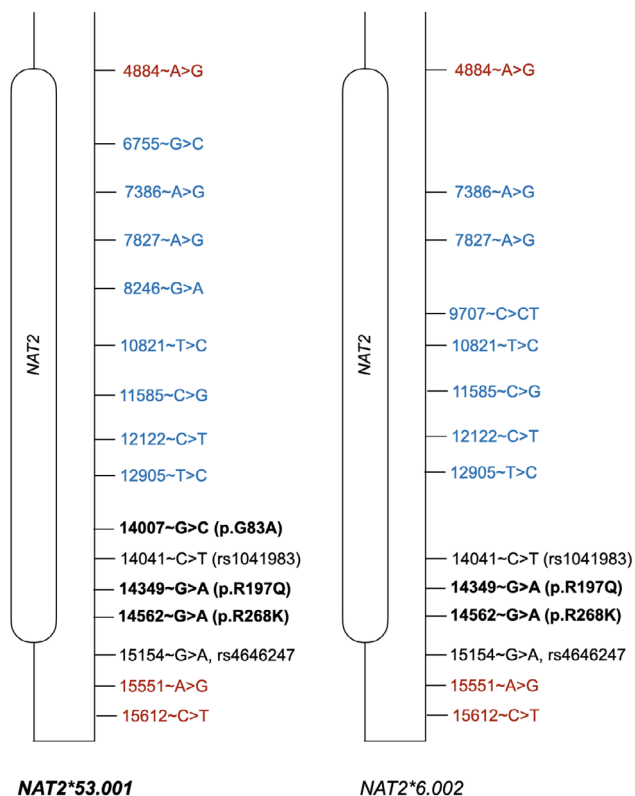


Figure 2 Depiction of a novel NAT2 star allele (*53.001) in HG03270—among the Esan in Nigeria (ESN) participants in the 1000 Genomes Project datasets analyzed in the study. HG03270 (Mother) has NAT2*6.002 as the second NAT2 haplotype and they are part of the NG65 Coriell trio. The father in the trio (HG03271) has NAT2*5.002/*6.002. The child (HG03272) in the trio inherited the novel star allele (*53.001) from the mother and *6.002 from the father. The variants in blue font are in intron 1 and are not captured in the current PharmVar NAT2 nomenclature file. The variants in red font are outside of the region being used for allele definitions (<https://www.pharmvar.org/gene/NAT2>).

of these two deletions are speculated to have been impacted by selection pressures such as diet, exposure to toxins and susceptibility to disease.⁴⁹ Generally, *GSTM1**0 is less frequent compared with *GSTT1**0 in SSA, which partly accounts for the low frequency of the *GSTM1*-*GSTT1* double homozygous deletion observed across SSA (7%) and the other biogeographical groups. However, relatively higher prevalences (11–14%) have been previously observed in West African populations.^{47,50} The *GSTM1*-*GSTT1* double null genotype has been shown to be associated with AT-DILI,¹¹ and it is also associated with an increased risk of hepatocellular carcinoma.⁵¹ Further studies on the influence of the *GSTM1*-*GSTT1* null deletion on anti-TB drug response in underrepresented SSA population groups are needed; given the considerable occurrence in SSA. For *GSTP1*, the reference haplotype (*A) has been associated with hepatotoxicity in individuals being administered with anti-TB drugs⁵² and increased risk of ADRs to anti-cancer medication.⁵³ However, in SSA most of the haplotypes estimated were defined by rs1695 (*GSTP1**B). The rs1695 frequency ranged from 34.7% (BRN) to 61.0% (MAL), which is relatively higher than

that reported in the Kassogue et al. study (12.1–49.8%) across Malian participants.⁴⁷

Among the *CYP2E1* variants, rs2070676 (present in the definition of *1B and other haplotypes), rs2070673, and rs2070672 (defining *7C) were observed at high frequencies (>16%) in the SSA populations. Additionally, of the nonsynonymous variants used to define *CYP2E1* star alleles, rs6413419 (*CYP2E1**4) was mostly African-specific. The rs6413419 variant is associated with normal activity,³² but was predicted to be deleterious by PolyPhen, CADD, LRT, and PROVEAN in this study. Many of the nonsynonymous 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 harboring potentially novel *CYP2E1* star alleles. Instead, we inferred potential haplotypes harboring these variants using SHAPEIT4. Six different *CYP2E1* haplotypes had rs28969387 as one of the defining variants. This variant was predicted to be deleterious and has been observed to alter the *CYP2E1* enzyme activity.⁵⁴ Regarding *CYP2E1* structural variant analysis, SSA populations had a higher frequency of duplications compared with other global populations. The *CYP2E1* gene deletion was observed in one individual from the MSL and the SA populations in this study. Generally, *CYP2E1* structural variants and their association with anti-TB drug response is understudied. Given the lack of consistency in *CYP2E1* star allele definitions and lack of consensus on the pharmacogenetic effects of *CYP2E1* variants, more functional studies, especially in African populations where TB is prevalent, are critical in addressing these knowledge gaps.

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 characterizations and functional studies across underrepresented African populations for relevant drug-gene pairs. 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 several ambiguous diplotypes calls. This made it difficult to compare haplotypes called in this study with those previously published. 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. Where applicable, haplotypes were estimated using trio data or statistical phasing methods. Using data from multiple sources may have affected the accuracy of some of our frequency estimates; however, given the consistency of the read depth, variant calling, and filtering process, we believe our results are meaningful and sound. Fourthly, functional studies to confirm the clinical

impact of the variants predicted to be deleterious by *in silico* prediction tools, were beyond 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.

In conclusion, this study elucidates the pharmacogenomic variation in *NAT1*, *NAT2*, *GSTM1*, and *GSTT1*, *GSTP1*, and *CYP2E1* across various SSA populations, and in comparison with other global populations. Furthermore, this study highlights differences in the distribution of key star alleles—such as *NAT1**10, *NAT2**5, *NAT2**16, *GSTM1**0, and *GSTT1**0—between SSA populations and other global populations. Our findings on the predicted phenotype distributions warrant the need for further investigations on the potential for genotype-guided dosing regimens, especially for TB (INH and other first-line drugs), hypertension (hydralazine), and cancer drug therapy to mitigate the risk of ADRs such as DILI in SSA and globally. A vital next step would be to establish whether the predicted phenotype patterns correspond to observations in clinical-based pharmacogenomics studies involving the SSA populations represented in this study. This would aid in understanding the link of anti-TB drug response (especially for INH) to the associated activity of the star alleles identified in this study. 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. 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 characterize their sequence content and support their designation as novel alleles by PharmVar. Together with determining the possible impact that these novel haplotypes may have to INH response, this information will be critical in precision medicine strategies, such as developing effective pharmacogenetic testing panels, in Africa.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

We are grateful to colleagues in the H3Africa Consortium who generated the bulk of the datasets used in this study (<https://h3africa.org>). We acknowledge the New York Genome Center and partner institutions that generated the high-coverage 1000 Genomes Project datasets (NHGRI Grant 3UM1HG008901-03S1). We thank all the study participants for their generosity and invaluable contribution in making this research possible.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

FUNDING

This research is a substudy under a project titled “Approaches of GenOmics for Response to drugs in Africa-TB-Malaria” (AGORA-TM), which is funded through the Genomic Research Approach for Diversity and Optimising Therapeutics (GRADIENT) initiative, a partnership between GlaxoSmithKline (GSK), Novartis, and the South African Medical Research Council (SAMRC). T.V.B.M. was partially supported by the University of the Witwatersrand Postgraduate Merit Award. The whole-genome sequencing of participants in Human Heredity and Health in Africa (H3Africa) projects was supported by grants from the National Human Genome Research Institute, National Institutes of Health (NIH/NHGRI,

Grant U54HG003273) and the Wellcome Trust through the Science for Africa foundation. The AWI-Gen Collaborative Center is funded by the NIH/NHGRI (Grant U54HG006938) as part of the H3Africa Consortium. M.R. is a South African Research Chair in Genomics and Bioinformatics of African Populations hosted by the University of the Witwatersrand, funded by the Department of Science and Innovation (South Africa), and administered by National Research Foundation of South Africa (NRF). The primary work relating to DNA sample processing and WGS at the Cell Biology Research Lab is based on research supported by grant awards from the Strategic Health Innovation Partnerships (SHIP) Unit of the SAMRC, a grantee of the Bill and Melinda Gates Foundation, and the South African Research Chairs Initiative of the Department of Science and Technology, and National Research Foundation (NRF) of South Africa (84177). The opinions, findings, and conclusions or recommendations expressed in this manuscript are solely the responsibility of the authors and not necessarily to be attributed to GSK, Novartis, SAMRC, NRF, or the NIH/NHGRI.

AUTHOR CONTRIBUTIONS

T.V.B.M., H.O., M.P., C.T.T., S.H., and D.T. wrote the manuscript. T.V.B.M., H.O., and D.T. designed the research. T.V.B.M., D.T., and H.O. performed the research. T.V.B.M. analyzed the data. C.T.T. M.R. contributed new study datasets and materials.

CODE AVAILABILITY STATEMENT

The StellarPGx code can be accessed at <https://github.com/SBIMB/StellarPGx>. The code used for the statistical analysis can be accessed at <https://github.com/ThanDEKS/PGx-analysis-AGORA-TM>.

DATA AVAILABILITY STATEMENT

The 1000 Genomes data (www.internationalgenome.org/) and the SGDP data (<https://www.simonsfoundation.org/simons-genome-diversity-project/>) are publicly available. The SAHGP data, H3Africa data, and CBRL datasets used in this study are available from European Genome-Phenome Archive (<https://ega-archive.org/>) on application to the relevant Data Access Committees (EGADs: EGAD00001003791, EGAD00001004316, EGAD00001004393, EGAD00001004557, EGAD00001004505, EGAD00001004334, EGAD00001004220, EGAD00001004448, EGAD00001004533, EGAD00001006418, and EGAD00001007589).

© 2025 The Author(s). *Clinical Pharmacology & Therapeutics* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

1. World Health Organization. *Global Tuberculosis Report 2023*. 1–75 <<https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023>> (2023).
2. Haas, D.W. *et al.* Pharmacogenetics of between-individual variability in plasma clearance of bedaquiline and clofazimine in South Africa. *J. Infect. Dis.* **226**, 147–156 (2022).
3. Pietersen, E. *et al.* 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**, e0281097 (2023).
4. Yimer, G. *et al.* Evaluation of patterns of liver toxicity in patients on antiretroviral and anti-tuberculosis drugs: a prospective four arm observational study in Ethiopian patients. *PLoS One* **9**, e94271 (2014).
5. Mugusi, S. *et al.* Liver enzyme abnormalities and associated risk factors in HIV patients on efavirenz-based HAART with or without tuberculosis co-infection in Tanzania. *PLoS One* **7**, e40180 (2012).
6. Chen, R., Wang, J., Zhang, Y., Tang, S. & Zhan, S. Key factors of susceptibility to anti-tuberculosis drug-induced hepatotoxicity. *Arch. Toxicol.* **89**, 883–897 (2015).

7. Cavaco, M.J. *et al.* 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 (2022).
8. Yang, S., Hwang, S.J., Park, J.Y., Chung, E.K. & Lee, J.I. Association of genetic polymorphisms of CYP2E1, NAT2, GST and SLC01B1 with the risk of anti-tuberculosis drug-induced liver injury: a systematic review and meta-analysis. *BMJ Open* **9**, e027940 (2019).
9. Chan, S.L. *et al.* Association and clinical utility of NAT2 in the prediction of isoniazid-induced liver injury in Singaporean patients. *PLoS One* **12**, e0186200 (2017).
10. El-Jaick, K.B. *et al.* Homozygotes NAT2*5B slow acetylators are highly associated with hepatotoxicity induced by anti-tuberculosis drugs. *Mem. Inst. Oswaldo Cruz* **117**, e210328 (2022).
11. Chbili, C. *et al.* Glutathione-S-transferase genetic polymorphism and risk of hepatotoxicity to antitubercular drugs in a North-African population: a case-control study. *Gene* **809**, 146019 (2022).
12. Radouani, F. *et al.* A review of clinical pharmacogenetics studies in African populations. *Pers. Med.* **17**, 155–170 (2020).
13. Collins, K.S. *et al.* Genotype-guided hydralazine therapy. *Am. J. Nephrol.* **51**, 764–776 (2020).
14. Rostami, S., Tafvizi, F. & Manjili, H.R.K. High efficacy of tamoxifen-loaded L-lysine coated magnetic iron oxide nanoparticles in cell cycle arrest and anti-cancer activity for breast cancer therapy. *Bioimpacts* **12**, 301–313 (2022).
15. Butcher, N.J., Boukouvala, S., Sim, E. & Minchin, R.F. Pharmacogenetics of the arylamine N-acetyltransferases. *Pharmacogenomics J.* **2**, 30–42 (2002).
16. Byrsk-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
17. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
18. Ramsay, M. *et al.* H3Africa AWI-gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob. Health Epidemiol. Genomics* **1**, 1–13 (2016).
19. Choudhury, A. *et al.* Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* **8**, 1–12 (2017).
20. Twesigomwe, D. *et al.* StellarPGx: a Nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clin. Pharmacol. Ther.* **110**, 741–749 (2021).
21. Eggertsson, H.P. *et al.* GraphTyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
22. Townsend, D.M. & Tew, K.D. The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* **22**, 7369–7375 (2003).
23. Watson, M.A., Stewart, R.K., Smith, G.B.J., Massey, T.E. & Bell, D.A. Human glutathione S-transferase P1 polymorphisms: relationship to lung tissue enzyme activity and population frequency distribution. *Carcinogenesis* **19**, 275–280 (1998).
24. Delaneau, O., Coulonges, C. & Zagury, J.F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).
25. Sabbagh, A., Darlu, P., Crouau-Roy, B. & Poloni, E.S. Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. *PLoS One* **6**, e18507 (2011).
26. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).
27. Rodrigues, C.H.M., Pires, D.E.V. & Ascher, D.B. DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* **30**, 60–69 (2021).
28. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
29. Wang, D., Para, M.F., Koletar, S.L. & Sadee, W. Human N-acetyltransferase 1 (NAT1) *10 and *11 alleles increase protein expression via distinct mechanisms and associate with sulfamethoxazole-induced hypersensitivity. *Pharmacogenet. Genomics* **21**, 652–664 (2011).
30. McBride, O.W., Umeno, M., Gelboin, H.V. & Gonzalez, F.J. A Taq I polymorphism in the human P450HE1 gene on chromosome 10 (CYP2E). *Nucleic Acids Res.* **15**, 10071 (1987).
31. Hu, Y., Hakkola, J., Oscarson, M. & Ingelman-Sundberg, M. 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 (1999).
32. Fairbrother, K.S. *et al.* Detection and characterization of novel polymorphisms in the CYP2E1 gene. *Pharmacogenetics* **8**, 543–552 (1998).
33. Hayashi, S.-I., Watanabe, J. & Kawajiri, K. Genetic polymorphisms in the 5'-flanking region change transcriptional regulation of the human cytochrome P450IIE1 gene. *J. Biochem.* **110**, 559–565 (1991).
34. Persson, I. *et al.* Genetic polymorphism of cytochrome P4502E1 in a Swedish population. Relationship to incidence of lung cancer. *FEBS Lett.* **319**, 207–211 (1993).
35. Mehta, R., Ive, P., Evans, D. & Menezes, C.N. Treatment outcomes among patients admitted to hospital with antiretroviral and/or antituberculosis drug-induced liver injury. *S. Afr. Med. J.* **111**, 474–481 (2021).
36. Patin, E. *et al.* Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene. *Hum. Mutat.* **27**, 720 (2006).
37. Mortensen, H.M. *et al.* Characterization of genetic variation and natural selection at the Arylamine N-acetyltransferase genes in global human populations. *Pharmacogenomics* **12**, 1545–1558 (2011).
38. Sy, S.K.B. *et al.* 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 (2015).
39. Stocco, G. *et al.* Thiopurine metabolites variations during co-treatment with aminosalicylates for inflammatory bowel disease: effect of N-acetyl transferase polymorphisms. *World J. Gastroenterol.* **21**, 3571–3578 (2015).
40. Podgorná, E. *et al.* Variation in NAT2 acetylation phenotypes is associated with differences in food-producing subsistence modes and ecoregions in Africa. *BMC Evol. Biol.* **15**, 1–20 (2015).
41. McDonagh, E.M., Boukouvala, S., Akililu, E., Hein, D.W., Altman, R.B. & Klein, T.E. PharmGKB summary: very important pharmacogene information for N-acetyltransferase 2. *Pharmacogenet. Genomics* **24**, 409–425 (2014).
42. Du, H. *et al.* Slow N-acetyltransferase 2 genotype contributes to anti-tuberculosis drug-induced hepatotoxicity: a meta-analysis. *Mol. Biol. Rep.* **40**, 3591–3596 (2013).
43. Mthiyane, T. *et al.* N-acetyltransferase 2 genotypes among Zulu-speaking South Africans and isoniazid and N-acetyl-isoniazid pharmacokinetics during antituberculosis treatment. *Antimicrob. Agents Chemother.* **64**, e02376-19 (2020).
44. Ndong Sima, C.A.A., Othman, H. & Möller, M. Advancing pharmacogenetics research in Africa: the “Project Africa GRADIENT” initiative. *Drug Discov. Today* **29**, 103939 (2024).
45. Sanni, S., Sina, H. & Baba-Moussa, L. Genetic polymorphisms and toxicities of first-line antituberculosis drugs: systematic review of the literature. *J. Tuberc. Res.* **10**, 124–145 (2022).
46. Kasthurinaidu, S.P., Ramasamy, T., Ayyavoo, J., Dave, D.K. & Adroja, D.A. GST M1-T1 null allele frequency patterns in geographically assorted human populations: a phylogenetic approach. *PLoS One* **10**, e0118660 (2015).
47. Kassogue, Y. *et al.* Genetic polymorphism of drug metabolism enzymes (GSTM1, GSTT1 and GSTP1) in the healthy Malian population. *Mol. Biol. Rep.* **47**, 393–400 (2020).
48. Palma-Cano, L.E. *et al.* GSTT1 and GSTM1 null variants in Mestizo and Amerindian populations from northwestern Mexico and a literature review. *Genet. Mol. Biol.* **40**, 727–735 (2017).
49. Piacentini, S., Polimanti, R., Porreca, F., Martínez-Labarga, C., de Stefano, G.F. & Fuciarelli, M. GSTT1 and GSTM1 gene polymorphisms in European and African populations. *Mol. Biol. Rep.* **38**, 1225–1230 (2011).

50. Salem, A.H. *et al.* Genetic polymorphism of the glutathione S-transferase M1 and T1 genes in three distinct Arab populations. *Dis. Markers* **31**, 311–316 (2011).
51. Li, S. *et al.* GSTM1 and GSTT1 null genotype increase the risk of hepatocellular carcinoma: evidence based on 46 studies. *Cancer Cell Int.* **19**, 76 (2019).
52. Wu, S. *et al.* Genetic polymorphisms of glutathione S-transferase P1 (GSTP1) and the incidence of anti-tuberculosis drug-induced hepatotoxicity. *PLoS One* **11**, e0157478 (2016).
53. Tecza, K., Pamula-Pilat, J., Lanuszevska, J., Butkiewicz, D. & Grzybowska, E. Pharmacogenetics of toxicity of 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer patients. *Oncotarget* **9**, 9114–9136 (2018).
54. Preissner, S.C., Hoffmann, M.F., Preissner, R., Dunkel, M., Gewiess, A. & Preissner, S. Polymorphic cytochrome P450 enzymes (CYPs) and their role in personalized therapy. *PLoS One* **8**, eB2562 (2013).